Free energy computations by minimization of Kullback-Leibler divergence: an efficient adaptive biasing potential method for sparse representations

I. Bilionis

Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, USA

P.S. Koutsourelakis *

Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, USA

Abstract

The present paper proposes an adaptive biasing potential technique for the computation of free energy landscapes. It is motivated by statistical learning arguments and unifies the tasks of biasing the molecular dynamics to escape free energy wells and estimating the free energy function, under the same objective of minimizing the Kullback-Leibler divergence between appropriately selected densities. It offers rigorous convergence diagnostics even though history dependent, non-Markovian dynamics are employed. It makes use of a greedy optimization scheme in order to obtain sparse representations of the free energy function which can be particularly useful in multidimensional cases. It employs embarrassingly parallelizable sampling schemes that are based on adaptive Sequential Monte Carlo and can be readily coupled with legacy molecular dynamics simulators. The sequential nature of the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
14 OCT 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Free energy computations by minimization of Kullback-Leibler divergence: an efficient adaptive biasing potential method for sparse representations	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Cornell University,Center for Applied Mathematics,Ithaca,NY,14853**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The present paper proposes an adaptive biasing potential technique for the computation of free energy landscapes. It is motivated by statistical learning arguments and unifies the tasks of biasing the molecular dynamics to escape free energy wells and estimating the free energy function, under the same objective of minimizing the Kullback-Leibler divergence between appropriately selected densities. It offers rigorous convergence diagnostics even though history dependent, non-Markovian dynamics are employed. It makes use of a greedy optimization scheme in order to obtain sparse representations of the free energy function which can be particularly useful in multidimensional cases. It employs embarrassingly parallelizable sampling schemes that are based on adaptive Sequential Monte Carlo and can be readily coupled with legacy molecular dynamics simulators. The sequential nature of the learning and sampling scheme enables the efficient calculation of free energy functions parametrized by the temperature. The characteristics and capabilities of the proposed method are demonstrated in three numerical examples.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**58**	

learning and sampling scheme enables the efficient calculation of free energy functions parametrized by the temperature. The characteristics and capabilities of the proposed method are demonstrated in three numerical examples.

Key words: free energy computations, adaptive biasing potential, Sequential Monte Carlo, atomistic simulations, statistical learning
PACS:

1 Introduction

Free energy is a central concept in thermodynamics and in the study of several systems in biology, chemistry and physics [11]. It represents a rigorous way to coarse-grain systems consisting of very large numbers of atomistic degrees of freedom, to probe states not accessible experimentally, to characterize global changes as well as investigate relative stabilities. In most applications, a brute-force computation based on sampling the atomistic positions is impractical or infeasible as the free energy barriers to overcome are so large that the system remains trapped in metastable free energy sets [56,61,11,77].

Equilibrium techniques for computing free energy surfaces such as Thermodynamic Integration [35], Weighted Histogram Analysis Method (WHAM, [39,67]), Adaptive Integration [71,26], Multistate Bennett Acceptance Ratio (MBAR, [66]) require the simulation of very long atomistic trajectories in order to achieve equilibrium. Furthermore, sampling along these paths correctly

* Corresponding Author. Tel: 607-254-5441
Email addresses: ib227@cornell.edu (I. Bilionis), pk285@cornell.edu (P.S. Koutsourelakis).

might necessitate advanced and quite involved techniques [13]. Techniques based on non-equilibrium path sampling [31,32,28,30] lack adaptivity and require the user to specify a particular path on the reaction coordinate space connecting two energetically important free energy regions, which can be non-trivial a task. More recently proposed adaptive biasing potential [3,76,41,2,24] and adaptive biasing force [17,16,29] techniques are capable of dynamically utilizing information obtained from the atomistic trajectories to bias the current dynamics in order to facilitate the escape from metastable sets [43]. They are able to automatically discover important regions of the reaction coordinate space. Since they rely on history-dependent, non-Markovian dynamics, it is not a priori clear, and in which sense, the system reaches a stationary state. Some work along these lines has been done in the context of the adaptive biasing force method in [44], for Langevin-type systems in [6] and in [53,43].

We propose an adaptive biasing potential technique where the two tasks of biasing the dynamics and estimating the free energy landscape are unified under the same objective of minimizing the Kullback-Leibler divergence between appropriately selected distributions on the extended space that includes atomic coordinates and the collective variables [51,52]. This framework provides a natural way for selecting the basis functions used in the approximation of the free energy and obtaining sparse representations which is critical when multi-dimensional collective variables are used. It allows the analyst to utilize and correct any prior information on the free energy landscape and provides an efficient manner of obtaining good estimates at various temperatures. The scheme proposed is embarrassingly parallelizable and relies on adaptive Sequential Monte Carlo procedures which enable efficient sampling from the high-dimensional and potentially multi-modal distributions of interest.

The structure of the rest of the paper is as follows. In the beginning of Section 2 we motivate our method for the alchemical case arriving at the identification of three (interconnected) problems: the selection of a parametrization of the free energy function, the choice of a distance metric in the space of probability densities and an optimization scheme to minimize this distance between two appropriately selected densities. Section 2.1 discusses the suitability and advantages of the Kullback-Leibler divergence. Section 2.2 deals with the optimization strategy employed and the use of a stochastic approximation scheme that guarantees convergence under weak conditions. Section 2.3 discusses a suboptimal strategy for the successive resolution of the free energy landscape by progressive addition of basis functions. Section 2.4 is concerned with an adaptive Sequential Monte Carlo scheme for the estimation of the expectations involved. Section 3 demonstrates the advantages of the proposed method for two important extensions: the reaction coordinate case, and the calculation of the free energy function as a function of temperature. Finally Section 4 contains results that illustrate the capabilities of the method with numerical results from three test cases.

2 Methodology - A statistical learning approach for adaptively calculating free energies

For clarity of the presentation, we will first introduce our method for the so-called alchemical case and generalize it later for the reaction coordinate case. Consider a molecular system with (generalized) coordinates $\boldsymbol{q} \in \mathcal{M} \subset \mathbb{R}^n$ following a Boltzmann-like distribution which in turn depends on some

parameters $\mathbf{z} \in \mathcal{D} \subset \mathbb{R}^d$ (in general $d << n$):

$$p(\boldsymbol{q}|\mathbf{z}) \propto \exp\left(-\beta V(\boldsymbol{q};\mathbf{z})\right), \tag{1}$$

where $V(\boldsymbol{q};\mathbf{z})$ is the potential energy of the system and β plays the role of inverse temperature. The free energy $A(\mathbf{z})$ is defined, up to an additive constant, as:

$$A(\mathbf{z}) = -\beta^{-1} \log \int_{\mathcal{M}} \exp\left(-\beta V(\boldsymbol{q};\mathbf{z})\right) d\boldsymbol{q}. \tag{2}$$

In general, the paperameters $\mathbf{z}$ provide a coarse-grained description of the molecular system and $A(\mathbf{z})$ concisely summarizes the system's behavior with respect to those variables. Our goal is to compute the free energy function $A(\mathbf{z})$ over the whole domain $\mathcal{D}$.

Let $\hat{A}(\mathbf{z};\boldsymbol{\theta})$ be an estimate of $A(\mathbf{z})$ parametrized by $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$. This parametrization will be made precise in the sequel. We define a joint probability distribution on the (generalized) coordinates $\boldsymbol{q}$ and the parameters $\mathbf{z}$ as:

$$p(\mathbf{q},\mathbf{z} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} 1_{\mathcal{D}}(\mathbf{z}) e^{-\beta\left(V(\mathbf{q},\mathbf{z}) - \hat{A}(\mathbf{z};\boldsymbol{\theta})\right)}, \tag{3}$$

where $1_{\mathcal{D}}(\mathbf{z})$ is the indicator function on $\mathcal{D}$ and $Z(\boldsymbol{\theta})$ is the normalization constant, i.e.:

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{D}\times\mathcal{M}} e^{-\beta\left(V(\mathbf{q},\mathbf{z}) - \hat{A}(\mathbf{z};\boldsymbol{\theta})\right)} d\boldsymbol{q} d\mathbf{z}. \tag{4}$$

It is easy to verify that the marginal density of the parameters $\mathbf{z} \in \mathcal{D}$ is given by:

$$\begin{aligned} p(\mathbf{z} \mid \boldsymbol{\theta}) &= \textstyle\int_{\mathcal{M}} p(\mathbf{q},\mathbf{z} \mid \boldsymbol{\theta})\, d\mathbf{q} \\ &= \tfrac{1}{Z(\boldsymbol{\theta})} 1_{\mathcal{D}}(\mathbf{z}) e^{-\beta\left(A(\mathbf{z}) - \hat{A}(\mathbf{z};\boldsymbol{\theta})\right)}. \end{aligned} \tag{5}$$

The key property of $p(\mathbf{z} \mid \boldsymbol{\theta})$ is that it reduces to the uniform distribution on $\mathcal{D}$ if and only if the free energy estimate is exact (up to an additive constant),

i.e. $\hat{A}(\mathbf{z};\boldsymbol{\theta}) = A(\mathbf{z}),\ \mathbf{z} \in \mathcal{D}$. As a result a natural strategy to estimate $A(\mathbf{z})$ is by minimizing a distance metric between $p(\mathbf{z} \mid \boldsymbol{\theta})$ and the uniform distribution over $\mathcal{D}$.

To make things mathematically precise, let $\mathcal{P}$ denote the the set of all probability densities on $\mathcal{D}$ with respect to $\mathbf{z}$, and:

$$\pi(\mathbf{z}) = 1_{\mathcal{D}}(\mathbf{z}) \frac{1}{\mid \mathcal{D} \mid} \tag{6}$$

the uniform density on $\mathcal{D}$ (whose volume is denoted by $\mid \mathcal{D} \mid$). Our approach to the free energy estimation problem consists of three key ingredients:

(1) the selection of a parameterization for $\hat{A}(\mathbf{z};\theta)$,

(2) the choice of a distance metric $\mathrm{d} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$

(3) a procedure for the solution of the minimization problem

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{d}\left(\pi(\mathbf{z}), p(\mathbf{z}|\boldsymbol{\theta})\right).$$

With regards to the first ingredient, we adopt representations inspired by kernel regression expansions. Kernel regression models have been proven successful for functional approximations in high-dimensional cases where d is in the order of 10 or 100 [72,73]. The unknown function is selected from a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ induced by a positive, semi-definite kernel $K(\cdot,\cdot)$. In particular, the approximate free energy function $\hat{A}(\mathbf{z};\boldsymbol{\theta})$ is expressed as:

$$\hat{A}(\mathbf{z};\boldsymbol{\theta}) = \sum_{j=1}^{k} \theta_j K(\mathbf{z}, \mathbf{z}_j; \boldsymbol{\tau}_j) =: \sum_{j=1}^{k} \theta_j K_j(\mathbf{z}), \quad \mathbf{z} \in \mathcal{D} \tag{7}$$

where $\mathbf{z}_j$ are points in $\mathcal{D}$ and $\boldsymbol{\tau}_j$ kernel parameters whose role is described in the sequel. In order to fix the additive constant, we select a point $\mathbf{z}_0 \in \mathcal{D}$ such

that $\hat{A}(\mathbf{z}_0;\boldsymbol{\theta}) = 0$ [1]. In relevant literature different types of kernel functions have been used such as thin plate splines, multiquadrics or Gaussians. While all these functions can be employed in the framework presented, we focus our presentation on Gaussian kernels which also have an intuitive parametrization with regards to the *scale of variability* of $\hat{A}$ as quantified by the bandwidth parameters $\boldsymbol{\tau}_j = \{\tau_{j,l}\}_{l=1}^d$ in each dimension:

$$K_j(\mathbf{z}) = K(\mathbf{z}, \mathbf{z}_j; \boldsymbol{\tau}_j) = \exp\{-\sum_{l=1}^{d} \tau_{j,l}(z_l - z_{j,l})^2\}. \tag{8}$$

Gaussian kernels in the context of free energy approximations have also been used in [41,51,24]. In Section 2.3, we discuss a way of adaptively determining the cardinality of the expansion k, as well as the precise form of K_j (i.e.. $\mathbf{z}_j$ and $\boldsymbol{\tau}_j$).

With regards to second ingredient, we employ the Kullback-Leibler divergence as a measure of distance between probability distributions. As discussed in Section 2.1, this choice possesses computational and theoretical advantages and leads to a convex optimization problem. Finally with regards to the third ingredient, we propose a stochastic gradient descent scheme as discussed in detail in Section 2.2. The latter involves a stochastic approximation scheme and a sampler for the estimation of expectations involved which is discussed in Section 2.4. The algorithmic steps are summarized in Algorithm 2.

[1] This is always possible by changing the kernels in Equation (7) to $K'_j(\mathbf{z}, \mathbf{z}_j) = K_j(\mathbf{z}) - K_j(\mathbf{z}_0, \mathbf{z}_j)$.

2.1 Choice of the metric

We propose employing the Kullback-Leibler (KL) divergence $\mathrm{KL}(\pi(\mathbf{z}) \parallel p(\mathbf{z} \mid \boldsymbol{\theta}))$ [14]:

$$\mathrm{KL}(\pi \parallel p) = \int_{\mathcal{D}} \pi(\mathbf{z}) \log \frac{\pi(\mathbf{z})}{p(\mathbf{z} \mid \boldsymbol{\theta})} \, d\mathbf{z}. \tag{9}$$

It is always non-negative and becomes zero if and only if $\pi(\mathbf{z}) \equiv p(\mathbf{z} \mid \boldsymbol{\theta})$ or equivalently $\hat{A}(\mathbf{z}; \boldsymbol{\theta}) = A(\mathbf{z}), \ \mathbf{z} \in \mathcal{D}$ [2]. Despite the fact that it is not a metric in the mathematical sense, it is frequently used as a measure of the distance between two probability distributions. Furthermore the KL-divergence provides upper bounds to other commonly used metrics such as the *Hellinger distance* defined by:

$$\mathrm{H}(\pi \parallel p) = \left(\int_{\mathcal{D}} \left(\sqrt{p(\mathbf{z}|\boldsymbol{\theta})} - \sqrt{\pi(\mathbf{z})} \right)^2 \, d\mathbf{z} \right)^{1/2},$$

as well as the *total variation distance*: [3]

$$\mathrm{V}(\pi \parallel p) = \sup_{B \in \mathcal{B}(\mathcal{D})} \left| \int_B \left(p(\mathbf{z}|\boldsymbol{\theta}) - \pi(\mathbf{z}) \right) \, d\mathbf{z} \right|.$$

Le Cam's inequalities as well as Lemma 2.4 of [74] imply that:

$$\mathrm{V}^2(\pi \parallel p) \leq \mathrm{H}^2(\pi \parallel p) \leq KL(\pi \parallel p). \tag{10}$$

Hence an minimization of the KL-divergence provides good approximations of the free energy surface with respect to these two genuine distances as well.

Since:

$$\mathrm{KL}(\pi \parallel p) = -\log \mid \mathcal{D} \mid - \int \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta}) d\mathbf{z},$$

[2] As already mentioned, of interest are free-energy *differences* and therefore perturbations of $A(\mathbf{z})$ or $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$ by a constant are ignored.

[3] We denote here with $\mathcal{B}(\mathcal{D})$ the set of the Borel sets of $\mathbb{R}^d$ that are subsets of $\mathcal{D}$, i.e. the set of all Lebesgue measurable subsets of $\mathcal{D}$.

the aforementioned formulation offers a clear strategy for estimating the free energy by *minimizing* the following form with respect to $\boldsymbol{\theta}$:

$$I(\boldsymbol{\theta}) = -\int \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta}) d\mathbf{z}. \tag{11}$$

The KL-divergence is always non-negative, so the objective function $I(\boldsymbol{\theta})$ has a lower bound:

$$I(\boldsymbol{\theta}) \geq \log \mid \mathcal{D} \mid . \tag{12}$$

This lower bound can be readily calculated and used to monitor convergence as well as the quality of the approximation obtained.

Notice that $I(\boldsymbol{\theta})$ depends on the unknown free energy $A(\mathbf{z})$ explicitly (from Equation (5)):

$$\begin{aligned} I(\boldsymbol{\theta}) &= -\int \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta}) d\mathbf{z} \\ &= \beta \int \pi(\mathbf{z}) \left(A(\mathbf{z}) - \hat{A}(\mathbf{z}; \boldsymbol{\theta}) \right) d\mathbf{z} + \log Z(\boldsymbol{\theta}). \end{aligned} \tag{13}$$

However, its gradient $\boldsymbol{J}(\boldsymbol{\theta}) = \frac{\partial I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ depends on $A(\mathbf{z})$ only through an expectation over $p(\mathbf{z}|\boldsymbol{\theta})$. In particular, by differentiating Equation (4) we obtain:

$$\frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_j} = \frac{1}{Z(\boldsymbol{\theta})} \frac{\partial Z(\boldsymbol{\theta})}{\partial \theta_j} = \beta E_{p(\mathbf{z}|\boldsymbol{\theta})}[K_j(\mathbf{z})] \tag{14}$$

and thus:

$$\begin{aligned} J_j(\boldsymbol{\theta}) &= \tfrac{\partial I(\boldsymbol{\theta})}{\partial \theta_j} \\ &= -\beta E_{\pi(\mathbf{z})} \left[\tfrac{\partial \hat{A}}{\partial \theta_j} \right] + \tfrac{\partial \log Z}{\partial \theta_j} \\ &= \beta \left(E_{p(\mathbf{z}|\boldsymbol{\theta})} \left[K_j(\mathbf{z}) \right] - E_{\pi(\mathbf{z})} \left[K_j(\mathbf{z}) \right] \right), \end{aligned} \tag{15}$$

where $E_{p(\mathbf{z}|\boldsymbol{\theta})}[\cdot]$ and $E_{\pi(\mathbf{z})}[\cdot]$ imply an expectation with regards to $p(\mathbf{z}|\boldsymbol{\theta})$ and $\pi(\mathbf{z})$ respectively. Given the unavailability of $p(\mathbf{z} \mid \boldsymbol{\theta})$ (since it depends on the unknown free energy $A(\mathbf{z})$), the expectations above can only be computed by

Monte Carlo sampling in the joint space $\mathcal{M} \times \mathcal{D}$ with respect to the joint density $p(\boldsymbol{q}, \mathbf{z} \mid \boldsymbol{\theta})$ (Equation (3)) that is known up to the normalization constant. The algorithmic scheme employed for the computation of these expectations is discussed in detail in Sections 2.2 and 2.4.

Finally, it is important to note that the Hessian of the objective function $\frac{\partial^2 I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ is proportional to the covariance between the kernels i.e.:

$$\begin{aligned}
\frac{\partial^2 I}{\partial \theta_j \partial \theta_l} &= \frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l} \\
&= -\frac{1}{Z^2(\boldsymbol{\theta})} \frac{\partial Z}{\partial \theta_j} \frac{\partial Z}{\partial \theta_l} + \frac{1}{Z(\boldsymbol{\theta})} \frac{\partial^2 Z}{\partial \theta_j \partial \theta_l} \\
&= \beta^2 E_{p(\mathbf{z}|\boldsymbol{\theta})} \left[(K_j(\mathbf{z}) - E_{p(\mathbf{z}|\boldsymbol{\theta})}[K_j(\mathbf{z})])(K_l(\mathbf{z}) - E_{p(\mathbf{z}|\boldsymbol{\theta})}[K_l(\mathbf{z})]) \right] \\
&= \beta^2 \operatorname{Cov}_{p(\mathbf{z}|\boldsymbol{\theta})}[K_j, K_l].
\end{aligned} \tag{16}$$

As long as the covariance matrix is positive definite, the objective function is convex with respect to $\boldsymbol{\theta}$ and there is a unique minimum. For the Gaussian kernels employed (Equation (8)) which have infinite support, this condition is satisfied except for degenerate cases of $p(\mathbf{z} \mid \boldsymbol{\theta})$. The formulation presented can also be interpreted by using arguments based on the well-known Expectation-Maximization algorithm (EM, [23]) as discussed in appendix B where potential Bayesian extensions are also presented.

2.2 Optimization with noisy gradients

We propose employing a gradient descent scheme in order to determine $\boldsymbol{\theta}$. It is noted that for similar computational problems as they appear for example in the context of maximum entropy estimation, more involved procedures such as Improved Iterative Scaling [4,22] and noisy conjugate gradients [65] have

been be employed. Second-order (quasi-)Newton-Raphson techniques are also possible although the unavoidable Monte Carlo noise in the computation of the Hessian (i.e. the covariance in Equation (16)) can destroy its positive definiteness.

Let $\boldsymbol{\theta}^k$ denote the vector of kernel amplitudes (Equation (7)) when k such kernels are used. Let also $\boldsymbol{\theta}^k_m$ denote the estimate of $\boldsymbol{\theta}^k$ after m iterations of the gradient descent algorithm. Then at the $(m+1)-$iteration, the following update equation could be used:

$$\boldsymbol{\theta}^k_{m+1} = \boldsymbol{\theta}^k_m - \lambda \boldsymbol{J}(\boldsymbol{\theta}^k_m) \tag{17}$$

where $\lambda > 0$ is the learning rate.

In general exact calculation of the gradient $\boldsymbol{J}(\boldsymbol{\theta})$ (Equation (15)) is impossible and one must resort to noisy, Monte Carlo estimates. The noise can impede convergence or even lead to a divergent scheme. For that purpose we propose employing a stochastic approximation variant of the Robbins & Monro scheme [62,9] in combination with an adaptive Sequential Monte Carlo sampler. The former ensures convergence with finite sample size and does not necessitate equilibrium samples from $p(\mathbf{z} \mid \boldsymbol{\theta})$. The latter produces estimators with lower variance as compared to standard Markov Chain Monte Carlo schemes and leads to accelerated convergence. It is noted that stochastic approximation schemes in the context of free energy estimation have previously been used in [46,47].

If $\hat{\boldsymbol{J}}(\boldsymbol{\theta}^k_m)$ denotes the Monte Carlo estimate of the gradient obtained with a finite sample size (the details of this estimator are discussed in section 2.4),

then at the m^{th} iteration we update $\boldsymbol{\theta}^k$ as follows:

$$\boldsymbol{\theta}^k_{m+1} = \boldsymbol{\theta}^k_m - \eta_m \hat{\boldsymbol{J}}(\boldsymbol{\theta}^k_m), \tag{18}$$

where $\{\eta_m\}$ is an appropriately chosen sequence of learning rates. According to [68] (page 106), the aforementioned scheme converges almost surely to the root $\tilde{\boldsymbol{\theta}}^k$ of $\boldsymbol{J}(\boldsymbol{\theta})$:

$$\boldsymbol{J}(\tilde{\boldsymbol{\theta}}^k) = 0,$$

if the following four conditions are satisfied:

C1) (Learning Rates): $\eta_m > 0, \eta_m \to 0, \sum_{m=0}^{\infty} \eta_m = \infty$ and $\sum_{m=0}^{\infty} \eta_m^2 < \infty$.

C2) (Search Direction): For some symmetric, positive definite matrix $\boldsymbol{B}$ and every $0 < \epsilon < 1$,

$$\inf_{\epsilon < \|\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k\| < 1/\epsilon} \left(\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k\right)^T \boldsymbol{B}\boldsymbol{J}(\boldsymbol{\theta}^k) > 0.$$

C3) (Mean-zero noise): The estimator $\hat{\boldsymbol{J}}(\boldsymbol{\theta}^k)$ of $\boldsymbol{J}(\boldsymbol{\theta}^k)$ is unbiased i.e.:

$$E\left[\hat{\boldsymbol{J}}(\boldsymbol{\theta}^k)\right] = \boldsymbol{J}(\boldsymbol{\theta}^k), \tag{19}$$

C4) (Growth and variance bounds):

$$E\left[\| \hat{\boldsymbol{J}}(\boldsymbol{\theta}) \|^2\right] \leq c\left(1 + \| \boldsymbol{\theta} \|^2\right). \tag{20}$$

In this work we use sequences of the form:

$$\eta_m = \frac{\eta}{(m + m_0)^{-\alpha}},$$

with $1/2 < \alpha \leq 1, \eta > 0$ and $m_0 \geq 0$, which ensures that condition (1) is satisfied. In all numerical examples of Section 4 we used $m_0 = 0$. In Appendix A, we prove that condition C2 holds in our case for $\boldsymbol{B} = \boldsymbol{I}$ i.e. the identity matrix. Condition C3 is a trivial property of the Sequential Monte Carlo sampler

as discussed in Section 2.4 [19]. Finally, the fourth condition C4 (for bounded kernel functions K_j (Equation (15))) is satisfied for Sequential Monte Carlo estimators under mild condition as it has been shown in [15] (Theorem 1) and [40] (Theorem 4). More recent works have increased the generality of these results [12,21] and provided asymptotic rates for the variance of the estimators as well.

In practical terms, Equation (18) implies computing a weighted average of the gradient's estimates at the current and previous iterations. By employing a decreasing sequence of weights, information from the earlier iterations gets discarded gradually and more emphasis is placed on the recent iterations. In practice the gradient descent was terminated when the following two conditions were met. Firstly, when all all the components of the gradient $\hat{\boldsymbol{J}}(\boldsymbol{\theta})$ had crossed zero at least once. Given the problem convexity, this indicated that further flucluations were the result of noise in the Monte Carlo estimators. Secondly, when the relative change in $\boldsymbol{\theta}$ was smaller than a prescribed tolerance i.e. $\frac{\|\boldsymbol{\theta}^k_{m+1} - \boldsymbol{\theta}^k_m\|}{\|\boldsymbol{\theta}^k_m\|} < 10^{-3}$.

The final number of iterations depends also on the number of kernels present in the approximation.

2.3 Kernel Selection

A critical objective in the proposed framework relates to the *sparseness* of the free energy approximation i.e. the cardinality k of the expansion in Equation (7). This is important in at least two ways. Firstly, because sparser representations can more clearly expose salient features of the free energy landscape, and

as a consequence, of the atomistic ensemble considered. Secondly, because they reduce the number of parameters $\boldsymbol{\theta}$ with respect to which the optimization problem needs to be solved (section 2.2). Given a vocabulary of potentially overcomplete basis functions and a prescribed k, the problem amounts to identifying those kernels (Equation (7)) that best approximate the true free energy surface i.e. minimize the KL divergence for $\boldsymbol{z} \in \mathcal{D}$ (Equation (9)). Given that the objective function $I(\boldsymbol{\theta})$ implicitly depends on the kernels selected, the aforementioned problem is equivalent to finding, amongst all k-sized combinations of kernels, the one that gives rise to the smallest $I(\boldsymbol{\theta})$. This obviously implies an excessive computational effort since the cumbersome optimization problem with respect to $\boldsymbol{\theta}$ would have to be solved for all possible $k-$sized combinations of basis functions.

For that purpose, we propose a *suboptimal* scheme that proceeds by adding a single kernel at the end of each optimization cycle with respect to $\boldsymbol{\theta}$ i.e. the cardinality k of the expansion (Equation (7)) increases by one. Similar greedy procedures have been successfully applied in maximum entropy problems [79,22,80]. Without loss of generality, one can consider a vocabulary of functions that consists of the Gaussian kernels discussed in Equation (8) which are parametrized by their locations $\mathbf{z}_j \in \mathcal{D}$ and bandwidths $\boldsymbol{\tau}_j = \{\tau_{j,l} \in \mathbb{R}^+\}_{l=1}^d$. Given k such kernels, let $\boldsymbol{\theta}^k = \{\theta_j^k\}_{j=1}^k$ (Equation (7)) denote the corresponding parameters that minimize $I(\boldsymbol{\theta})$ in Equation (11). *The goal of the ensuing scheme is to select* $\mathbf{z}_{k+1} \in \mathcal{D}$, $\boldsymbol{\tau}_{k+1} = \{\tau_{k+1,l} \in \mathbb{R}^+\}_{l=1}^d$ *of the next* $k+1$ *kernel*. Once this have been achieved, the corresponding $\boldsymbol{\theta}^k = \{\theta_j^{k+1}\}_{j=1}^{k+1}$ values are obtained by carrying out the optimization process

discussed in the previous section for the given set of $k+1$ kernels. Let also:

$$\hat{A}_k(\boldsymbol{z};\boldsymbol{\theta}^k) = \sum_{j=1}^{K} \theta_j^k K_j(\mathbf{z}) \quad \text{and} \quad \hat{A}_{k+1}(\boldsymbol{z};\boldsymbol{\theta}^{k+1}) = \sum_{j=1}^{k} \theta_j^{k+1} K_j(\mathbf{z}) \tag{21}$$

denote the associated free-energy approximations obtained using k and $k+1$ kernels respectively, and $p_k(\mathbf{q},\mathbf{z} \mid \boldsymbol{\theta}^k)$ and $p_{k+1}(\mathbf{q},\mathbf{z} \mid \boldsymbol{\theta}^{k+1})$ the corresponding densities in Equation (3). Note that in general $\theta_j^k \neq \theta_j^{k+1},\ j = 1,\ldots,k$ even though the first k kernels in the aforementioned expansions (Equation (21)) are identical. Since $\boldsymbol{\theta}^k$ minimize $I(\boldsymbol{\theta})$, the gradient of $I(\boldsymbol{\theta})$ must be zero at $\boldsymbol{\theta}^k$, i.e.:

$$\frac{\partial I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid_{\boldsymbol{\theta}^k} = 0 \rightarrow E_\pi\left[K_j(\mathbf{z})\right] = E_{p_k}\left[K_j(\mathbf{z})\right]. \tag{22}$$

Given that $\hat{A}_{k+1}(\boldsymbol{z};\boldsymbol{\theta}^{k+1})$ provides a better approximation to the true free energy (or at least just as good as $\hat{A}_k(\boldsymbol{z};\boldsymbol{\theta}^k)$), the improvement in terms of Kullback-Leibler divergence (Equation (9)), denoted by Δ_{k+1} can be assessed with:

$$\begin{aligned} \Delta_{k+1} &= \mathrm{KL}(\pi \parallel p_k) - \mathrm{KL}(\pi \parallel p_{k+1}) \\ &= \beta E_\pi\left[\hat{A}_{k+1}(\boldsymbol{z};\boldsymbol{\theta}^{k+1}) - \hat{A}_k(\boldsymbol{z};\boldsymbol{\theta}^k)\right] - \log\tfrac{Z(\boldsymbol{\theta}^{k+1})}{Z(\boldsymbol{\theta}^k)}. \end{aligned} \tag{23}$$

By employing Equations (21) and (22), it can be shown that [79]

$$\begin{aligned} \Delta_{k+1} &= \beta E_\pi\left[\hat{A}_{k+1}(\boldsymbol{z};\boldsymbol{\theta}^{k+1}) - \hat{A}_k(\boldsymbol{z};\boldsymbol{\theta}^k)\right] - \log\tfrac{Z(\boldsymbol{\theta}^{k+1})}{Z(\boldsymbol{\theta}^k)} \\ &= \beta E_{p_{k+1}}\left[\hat{A}_{k+1}(\boldsymbol{z};\boldsymbol{\theta}^{k+1})\right] - \beta E_{p_k}\left[\hat{A}_k(\boldsymbol{z};\boldsymbol{\theta}^k)\right] - \log\tfrac{Z(\boldsymbol{\theta}^{k+1})}{Z(\boldsymbol{\theta}^k)} \\ &= KL(p_{k+1} \parallel p_k) \geq 0. \end{aligned} \tag{24}$$

Formally one would need to solve an optimization problem with respect to $\boldsymbol{\theta}^{k+1}$ for all possible $K_{k+1}(\mathbf{z})$ in order to find the one that maximizes Δ_{k+1} or equivalently the gain in terms of the KL-divergence. In order to circumvent

this difficulty, we employ a second-order Taylor expansion of Δ_{k+1} detailed in [79]:

$$\Delta_{k+1} \approx \frac{1}{2\,\mathrm{Var}_{k,k+1}} \left(E_{\pi}\left[K_{k+1}(\mathbf{z})\right] - E_{p_k}\left[K_{k+1}(\mathbf{z})\right] \right)^2, \tag{25}$$

where $\mathrm{Var}_{k,k+1}$ is the conditional variance of $K_{k+1}(\mathbf{z})$ given $K_j(\mathbf{z})$ $j = 1, \ldots, k$ with respect to a distribution that lies between p_k and p_{k+1} in the sense of Kullback (page 48, [38]). We propose therefore measuring this KL-gain using the difference between the expected values of $K_{k+1}(\mathbf{z})$ in terms of the (target) uniform distribution π and the current approximation p_k. This effectively suggests augmenting our expansion with the kernel that locally maximizes the gradient of $I(\boldsymbol{\theta})$. Intuitively it implies incorporating the kernel function whose expected value with respect to the target, uniform distribution is worst approximated by the current density p_k. In practical terms, and given the parametrization of the Gaussian kernels employed, this amounts to finding the location and bandwidth parameters $\left(\mathbf{z}^*_{k+1}, \boldsymbol{\tau}^*_{k+1} = \{\tau^*_{k+1}\}_{l=1}^{d}\right)$ (over the range of allowable values) so that:

$$(\mathbf{z}^*_{k+1}, \boldsymbol{\tau}^*_{k+1}) = \arg \max_{(\mathbf{z}_{k+1}, \boldsymbol{\tau}_{k+1})} \left(E_{p_k}\left[K(\mathbf{z}; \mathbf{z}_{k+1}, \boldsymbol{\tau}_{k+1})\right] - E_{\pi}\left[K(\mathbf{z}; \mathbf{z}_{k+1}, \boldsymbol{\tau}_{k+1})\right] \right)^2. \tag{26}$$

The solution of this simple, low-dimensional optimization problem can be carried out using any standard global/local technique. More details on the methodology adopted in the examples examined are contained in section 4. It is noted that the expectation with respect to p_k in Equation (26) is approximated using the Sequential Monte Carlo samplers discussed in section 2.4. Furthermore regularization terms can be added to the objective function of Equation (26) in order to promote lower or larger bandwidth kernels (see also appendix B). Naturally, the same formulation can be applied with any type of kernel or overcomplete basis employed (e.g. wavelets). The proposed strategy

promotes sparseness and computational efficiency while offering a progressive resolution of the free energy landscape that naturally involves kernels that carry most of the information in the first steps and successive unveiling of finer details (see the first example of section 4).

It is finally noted that once the next kernel has been selected and the optimization has been carried out, the KL-gain Δ_{k+1} (Equation (23)), offers a natural metric for monitoring convergence. The expectation with respect to the uniform can in general be calculated analytically whereas the ratio of normalizing constants $\log \frac{Z(\boldsymbol{\theta}^{k+1})}{Z(\boldsymbol{\theta}^{k})}$ (Equation (3)) is a direct output of the Sequential Monte Carlo sampling that is used to sample from the augmented densities and is discussed in the next section.

2.4 Adaptive Sequential Monte Carlo

The learning scheme proposed relies on efficient computations of the gradient appearing in Equation (15). This depends on expectations with respect to $p(\mathbf{z} \mid \boldsymbol{\theta})$ (Equation (5)) which are not available analytically since the actual free energy $A(\mathbf{z})$ is unknown. We resort to a Sequential Monte Carlo (SMC) scheme that draws samples from the joint density $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ in Equation (3) which involves the atomic degrees of freedom $\mathbf{q}$. It is noted however that convergence of the stochatic approximation algorithm discussed previously is guaranteed even with the most basic MCMC sampler. It is nevertheless important to have a sampling scheme that mixes well and reduces the bias in the learning. In addition, the SMC schemes proposed readily enable sampling from a sequence of distributions that is advantageous in obtaining several free energy landscapes (i.e. parametrized by the temperature) as illustrated in

section 3.2.

SMC samplers [20,18] represent a parallelizable strategy that combine the advantages of MCMC and Importance Sampling, resulting in lower variance estimators [49,34,33,70]. We propose novel extensions that allow the algorithm to automatically adapt to the difficulties of the target density, while retaining the ability to interact seamlessly with legacy, molecular dynamics simulators.

The proposed SMC schemes offer a flexible framework for sampling from a *sequence of unormalized probability distributions* and are therefore highly suited for the dynamic setting of the problem at hand where the target density $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ changes with $\boldsymbol{\theta}$. For a given $\boldsymbol{\theta}$, they approximate $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ with a set of N random samples (or *particles/replicas*) $\{\mathbf{q}^{(i)}, \mathbf{z}^{(i)}\}_{i=1}^{N}$, which are updated using a combination of *importance sampling*, *resampling* and MCMC-based *rejuvenation* mechanisms [19]. Each of these particles/replicas is associated with an *importance weight* $w^{(i)}$. The weights are updated sequentially along with the particle/replica locations in order to provide a particulate approximation:

$$p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}) \approx \sum_{i=1}^{N} W^{(i)} \, \delta_{\mathbf{q}^{(i)}}(\mathbf{q}) \delta_{\mathbf{z}^{(i)}}(\mathbf{z}), \tag{27}$$

where $W^{(i)} = w^{(i)} / \sum_{j=1}^{N} w^{(j)}$ are the normalized weights and $\delta_{\mathbf{z}^{(i)}}(\cdot)$ is the Dirac function centered at $\mathbf{z}^{(i)}$. These particles/replicas and weights can be used to estimate expectations of any $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$-integrable function which converge *almost surely* as $N \to \infty$ [18,12]. In particular for Equation (15):

$$\sum_{i=1}^{N} W^{(i)} \, K_j(\mathbf{z}^{(i)}) \to \int K_j(\mathbf{z}) \; p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}) \; d\mathbf{q} d\mathbf{z} = E_{p(\mathbf{z}|\boldsymbol{\theta})} \left[K_j(\mathbf{z}) \right]. \tag{28}$$

The proposed SMC algorithms will be used iteratively, after each step of the gradient descent algorithm. Given two successive estimates $\boldsymbol{\theta}_m^k$ and $\boldsymbol{\theta}_{m+1}^k$

(Equation (18)) and a particulate approximation of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_m^k)$, the goal is to obtain new samples from $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{m+1}^k)$ (Algorithm 2) and compute the new expectations in Equation (15) based on Equation (28). The quality of the Monte Carlo estimates in Equation (28) depends on the proximity of the distributions $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_m^k)$ and $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{m+1}^k)$. We propose building a path of intermediate, unormalized distributions that will bridge this gap based on Equation (3) [4]:

$$\begin{aligned} \pi_\gamma(\mathbf{q}, \mathbf{z}) &= p(\mathbf{q}, \mathbf{z} \mid (1-\gamma)\boldsymbol{\theta}_m^k + \gamma \boldsymbol{\theta}_{m+1}^k) \\ &= \exp\left\{-\beta\left(V(\mathbf{q}, \mathbf{z}) - \hat{A}(\mathbf{z}; \boldsymbol{\theta}_\gamma)\right)\right\}, \quad \gamma \in [0, 1], \end{aligned} \tag{29}$$

where

$$\boldsymbol{\theta}_\gamma = (1-\gamma)\boldsymbol{\theta}_m^k + \gamma \boldsymbol{\theta}_{m+1}^k. \tag{30}$$

Clearly for $\gamma = 0$ one recovers $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_m^k)$ and for $\gamma = 1$, $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{m+1}^k)$. The role of these auxiliary distributions is to provide a smooth transition path where importance sampling can be efficiently applied. Naturally, the more intermediate distributions are considered along this path, the higher the accuracy of the final estimates, but also the higher the computational cost. On the other hand too few intermediate distributions π_γ can adversely affect the overall accuracy of the approximation.

To that end we propose an *adaptive* SMC scheme that automatically determines the number of intermediate distributions needed [19,37]. In this process we are guided by the Effective Sample Size (ESS, [49]). In particular, let S be the total number of intermediate distributions (which is unknown a priori) and $\gamma_s,\ s = 1, 2, \ldots, S$ the associated bridging parameters such that

[4] subscripts k and m indicating the number of kernels and optimization iterations respectively have been dropped

$0 = \gamma_1 < \gamma_2 < \ldots < \gamma_S = 1$, which are also unknown a priori. Let also $\{(\mathbf{q}_s^{(i)}, \boldsymbol{z}_s^{(i)}),\ W_s^{(i)}\}_{i=1}^N$ denote the particulate approximation of π_{γ_s} defined as in Equation (29) for $\gamma = \gamma_s$. The Effective Sample Size of these particles/replicas is then defined as $\text{ESS}_s = 1/\sum_{i=1}^N (W_s^{(i)})^2$ and provides a measure of the population variance. One extreme, i.e. when $\text{ESS}_s = 1$, arises when a single replica has a unit normalized weight whereas the rest have zero weights and as a result provide no information. The other extreme, i.e. $\text{ESS}_s = N$, arises when all the replicas are equally informative and have equal weights $W_s^{(i)} = 1/N$.

If the next bridging distribution $\pi_{\gamma_{s+1}}$ is very similar to π_{γ_s} (ie. $\gamma_{s+1} \approx \gamma_s$), then ESS_{s+1} should not be that much different from ESS_s. On the other hand if that difference is pronounced then ESS_{s+1} could drop dramatically. Hence in determining the next auxiliary distribution, we define an acceptable reduction in the ESS, i.e. $\text{ESS}_{s+1} \geq \zeta\ \text{ESS}_s$ (where $\zeta < 1$ [5]) and prescribe γ_{s+1} (Equation (29)) accordingly.

The proposed adaptive SMC algorithm is summarized in Algorithm 1. The resampling component was carried out using a multinomial resampling scheme (i.e. the new population consisted of replicas drawn from the previous population with probability proportional to their weights) and was triggered for $\text{ESS}_{\min} = N/2$. It should be noted that unlike MCMC schemes, the particle/replica perturbations in the *Rejuvenation* step do not require that the $P_s(.,.)$ is *ergodic* [20]. It suffices that it is a π_{γ_s}-invariant kernel, which readily allows adaptively changing its parameters in order to achieve better mixing rates. In the examples presented a component-wise Metropolis-Hastings scheme ([5]) was used to update $\boldsymbol{q}$ and $\boldsymbol{z}$ separately by employing a Metropolis-

[5] The value $\zeta = 0.95$ was used throughout this study.

Algorithm 1 Adaptive SMC

Require: $s = 1$ and $\gamma_1 = 0$ and a population $\{(\mathbf{q}_1^{(i)}, \boldsymbol{z}_1^{(i)}),\ w_1^{(i)}\}_{i=1}^N$ which approximate $\pi_{\gamma_1} \equiv p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_m^k)$ in Equation (29).

Ensure: The population $\{(\boldsymbol{\theta}_s^{(i)}, \boldsymbol{z}_s^{(i)}), w_s^{(i)}\}_{i=1}^N$ provides a particulate approximation of π_{γ_s} in the sense of Equations (27), (28).

while $\gamma_s < 1$ **do**

$s \leftarrow s + 1$

{**Reweighting-Importance Sampling**}

Let

$$
\begin{aligned}
w_s^{(i)}(\gamma_s) &= w_{s-1}^{(i)} \ \frac{\pi_{\gamma_s}(\mathbf{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)})}{\pi_{\gamma_{s-1}}(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)})} \\
&= w_{s-1}^{(i)} \ \frac{\exp\left\{-\beta(V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) - \hat{A}(\boldsymbol{z}_{s-1}^{(i)}; \boldsymbol{\theta}_{\gamma_s})\right\}}{\exp\left\{-\beta(V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) - \hat{A}(\boldsymbol{z}_{s-1}^{(i)}; \boldsymbol{\theta}_{\gamma_{s-1}})\right\}} \\
&= w_{s-1}^{(i)} \ \exp\left\{-\beta(\hat{A}(\boldsymbol{z}_{s-1}^{(i)}); (\gamma_s - \gamma_{s-1})(\boldsymbol{\theta}_{m+1}^k - \boldsymbol{\theta}_m^k)\right\},
\end{aligned}
\tag{31}
$$

be the *updated* weights as a function of γ_s. Determine $\gamma_s \in (\gamma_{s-1}, 1]$ so that $\mathrm{ESS}_s = \zeta\, \mathrm{ESS}_{s-1}$.

{**Resampling**}

if $\mathrm{ESS}_s \leq \mathrm{ESS}_{\min}$ **then**

Resample

end if

{**Rejuvenation**}

Use an MCMC kernel $P_s\left((\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}), (\boldsymbol{q}_s^{(i)}, \boldsymbol{z}_s^{(i)})\right)$ that leaves π_{γ_s} invariant to perturb each replica $(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) \rightarrow (\boldsymbol{q}_s^{(i)}, \boldsymbol{z}_s^{(i)})$.

end while

Adjusted Langevin Algorithm (MALA) for each set of coordinates [63]. The $P_s(.,.)$ is therefore defined implicitely by the proposal density and the accept/reject step. Given $\left(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}\right)$, the proposals consist of:

- Updating $\boldsymbol{q}_{s-1}^{(i)} \rightarrow \boldsymbol{q}_{s}^{(i)}$:

$$\begin{aligned} \boldsymbol{q}_{s}^{(i)} - \boldsymbol{q}_{s-1}^{(i)} &= \tfrac{\Delta t_q}{2} \nabla_{\mathbf{q}} \log \pi_{\gamma_s}(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) + \sqrt{\Delta t_q}\boldsymbol{r}_q \\ &= -\tfrac{\beta \Delta t_q}{2} \nabla_{\mathbf{q}} V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) + \sqrt{\Delta t_q}\boldsymbol{r}_q. \end{aligned} \tag{32}$$

- Updating $\boldsymbol{z}_{s-1}^{(i)} \rightarrow \boldsymbol{z}_{s}^{(i)}$:

$$\begin{aligned} \boldsymbol{z}_{s}^{(i)} - \boldsymbol{z}_{s-1}^{(i)} &= \tfrac{\Delta t_z}{2} \nabla_{\mathbf{z}} \log \pi_{\gamma_s}(\boldsymbol{q}_{s}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) + \sqrt{\Delta t_z}\boldsymbol{r}_z \\ &= -\tfrac{\beta \Delta t_z}{2} \left(\nabla_{\mathbf{z}} V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) - \nabla_{\mathbf{z}} \hat{A}(\boldsymbol{z}_{s-1}^{(i)}; \boldsymbol{\theta}_{\gamma_s}) \right) + \sqrt{\Delta t_q}\boldsymbol{r}_q. \end{aligned} \tag{33}$$

where $\boldsymbol{r}_q$ and $\boldsymbol{r}_z$ are i.i.d standard Gaussian vectors. A Metropolis-Hastings accept/reject step with respect to the target invariant density $\pi_{\gamma_s}(\cdot)$ was performed after each update which ensures π_{γ_s}-invariance. Two different time steps were used Δt_q and Δt_z for the $\mathbf{q}$ and $\mathbf{z}$ coordinates respectively. Their values were adjusted after each iteration s so as to retain an average acceptance ratio (over all replicas N) between 50% and 80% [64], simply by increasing/decreasing the current time step using a multiplication factor [6]. The variable time step ensured better mixing at the Rejuvenation step and contributed to the adaptivity of the algorithm. The theoretical requirements are satisfied whether one or more MALA time steps are performed in Equations (32) and (33). Naturally other molecular dynamics samplers can be employed which could potentially exhibit better mixing or fit more closely to the physics of the problem at hand [8].

It is also noted that the approximation of the free energy $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$, biases the potential of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ (Equation (3)) and allows the system to overcome free

[6] The multiplication factors we used in the numerical examples where 1.2 for increase and 0.7 for decrease.

energy barriers [53]. Finally we note that the estimates of the ratio of normalization constants Z_s/Z_{s-1} between two successive unnormalized densities $\pi_{\gamma_{s-1}}$ and π_{γ_s}can be obtained by averaging the unnormalized updated weights in Equation (31) as a direct consequence of the importance sampling identity:

$$\begin{aligned}\frac{Z_s}{Z_{s-1}} &= \frac{\int \pi_{\gamma_s}(\mathbf{q},\mathbf{z})\ d\mathbf{q}d\mathbf{z}}{\int \pi_{\gamma_{s-1}}(\mathbf{q},\mathbf{z})\ d\mathbf{q}d\mathbf{z}} \\ &= \int \frac{\pi_{\gamma_s}(\mathbf{q},\mathbf{z})}{\pi_{\gamma_{s-1}}(\mathbf{q},\mathbf{z})} \frac{\pi_{\gamma_{s-1}}(\mathbf{q},\mathbf{z})}{Z_{s-1}}\ d\mathbf{q}d\mathbf{z} \\ &\approx \sum_{i=1}^{N} W_{s-1}^{(i)} \frac{\pi_{\gamma_s}(\mathbf{q}_{s-1}^{(i)},\mathbf{z}_{s-1}^{(i)})}{\pi_{\gamma_{s-1}}(\mathbf{q}_{s-1}^{(i)},\mathbf{z}_{s-1}^{(i)})}.\end{aligned} \tag{34}$$

These estimators can be telescopically multiplied ([20,36]) in order to compute the ratio of normalization constants between any pair of distributions as required in Equation (23).

Given the preceding discussion in sections 2.2, 2.3 and 2.4, we summarize below the basic steps in the proposed free energy computation scheme: In the inner loop and for fixed k, gradient descent (Section 2.2) is performed which makes use of the adaptive SMC scheme (Section 2.4) in order to compute the expectations in the gradient. In the outer loop, the cardinality of the expansion k is increased by adding one kernel (i.e. $k \leftarrow k+1$) based on Equation (26). This is terminated when the KL gain (Equation (23)) does not exceed a prescribed tolerance. Algorithm 2 summarizes formally the procedure.

Remark: As mentioned in section 2.2 a sufficient condition for the convergence of the proposed stochastic approximation scheme is the *unbiasedness* of the gradient estimators. It is noted that theoretical and numerical results suggest that this condition is more stringent than necessary and can be relaxed [50,58,78,1,10]. While the unbiasedness of SMC particulate approximations is

Algorithm 2 Calculation of the free energy at a given temperature.

Require: $k = 0$, $\boldsymbol{\theta}^0 \equiv \mathbf{0}$ and a particulate approximation of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}^0)$ (Equation (3)) at the desired temperature β (see Remark below).

while true do

Calculate Δ_k based on Equation (23).

if $\Delta_k \leq$ tol **then**

Break the loop.

else

Add the optimal $(k+1)^{th}$ kernel based on Equation (26) and set $k \leftarrow k+1$.

repeat

Estimate gradient at $\boldsymbol{\theta}_m^k$ and calculate update $\boldsymbol{\theta}_{m+1}^k$ based on Equation (18).

Use adaptive SMC (section 2.4) to construct particulate approximation of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{m+1}^k)$ from $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_m^k)$.

until Convergence criteria are met.

end if

end while

ensured by the importance sampling step, it relies on an unbiased estimator of the initial distribution $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}^0)$. In order to achieve this we considered two schemes, one approximate (but nearly exact) and one exact. The first involved running a long MCMC at a very high temperature which ensured good mixing (the distribution of the reaction coordinates was effectively uniform). We subsampled the chain in order to make sure that the samples drawn were close to independent and assigned equal weights. Subsequently the SMC scheme was employed to adaptively reduce the temperature. In physical problems where the calculation of the temperature-dependent free energy surface was of inter-

est (see section 3.2), this did not impose any additional burden. Furthermore, and as long as, the first temperature was high enough the error introduced in terms of bias was negligible as the MCMC chain for all practical purposes had attained equilibrium. The second menthod which is exact, relied on running MCMC for a few steps (at the target temperature) and estimating the mean and covariance of the atomistic coordinates $\boldsymbol{q}$ (and reaction coordinates $\boldsymbol{z}$ is the alchemical case). We subsequntly performed importance sampling using as an importance sampling density multivariate Gaussians with the aforementioned moments and calculated weights based on their ratio with the density $p(\boldsymbol{q}, \boldsymbol{z}|\boldsymbol{\theta}_0)$. This ensured an *unbiased estimator* for $p(\boldsymbol{q}, \boldsymbol{z}|\boldsymbol{\theta}_0)$ albeit a very poor one (with large variance) as the importance sampling distribution was in general a poor approximation of $p(\boldsymbol{q}, \boldsymbol{z}|\boldsymbol{\theta}_0)$. Despite the different initializations both methods converged in the sample problems examined which is a testament to the power of the stochastic approximation. We also performed runs where particles were generated by running MCMC with $p(\boldsymbol{q}, \boldsymbol{z}|\boldsymbol{\theta}_0)$ as the target. Despite the bias in the estimator (resulting from lack of equilibrium) the method was still able to converge which coincides with the results mentioned above.

3 Extensions

The current section is devoted to extensions of the proposed algorithmic environment. In particular we discuss the reaction coordinate case that generalizes the applicability of the proposed method (Section 3.1). Section 3.2 is devoted to the calculation of the free energy landscape as a function of the temperature where it is shown that the sequential nature of the proposed methodology can

lead to significant computational advantages.

3.1 The reaction coordinate case

The proposed method was described for the alchemical case. However, it is straightforwardly generalized to cover also the general reaction coordinate case. Let $\boldsymbol{\xi} : \mathcal{M} \to \mathcal{D}$ be a function of the system coordinates $\boldsymbol{q}$. This function is called a reaction coordinate [45]. It is evident that $\boldsymbol{q}$ can be viewed in a probabilistic framework as a random variable and as a result:

$$\mathbf{z} = \boldsymbol{\xi}(\boldsymbol{q}), \tag{35}$$

is also a random variable. The probability distribution of $\mathbf{z}$ can be found by integrating out $\boldsymbol{q}$:

$$p(\mathbf{z} \mid \beta) = \int p(\boldsymbol{q})\delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z})d\boldsymbol{q} \propto \int \exp\left(-\beta V(\boldsymbol{q})\right)\delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z})d\boldsymbol{q}. \tag{36}$$

The free energy $A(\mathbf{z})$ with respect to the reaction coordinate $\boldsymbol{\xi}(\boldsymbol{q})$ is defined to be the *effective potential* of $\mathbf{z} = \boldsymbol{\xi}(\boldsymbol{q})$, i.e.:

$$p(\mathbf{z}) \propto \exp\left(-\beta A(\mathbf{z})\right). \tag{37}$$

Combining these two equations we see that:

$$A(\mathbf{z}) = -\beta^{-1} \log \int \exp\left(-\beta V(\boldsymbol{q})\right)\delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z})d\boldsymbol{q}. \tag{38}$$

If $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$ is an estimate of $A(\mathbf{z})$, we define a new probability distribution over $\boldsymbol{q}$ as:

$$p(\boldsymbol{q}|\boldsymbol{\theta}) \propto 1_{\mathcal{D}}(\boldsymbol{\xi}(\boldsymbol{q})) \exp\left(-\beta(V(\boldsymbol{q}) - \hat{A}(\xi(\boldsymbol{q}); \boldsymbol{\theta}))\right). \tag{39}$$

It is straightforward to see that under this new distribution for $\boldsymbol{q}$, the pdf of

$\mathbf{z}$ becomes:

$$p(\mathbf{z}|\boldsymbol{\theta}) = \int p(\boldsymbol{q}|\boldsymbol{\theta})\delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z})d\boldsymbol{q} \propto 1_{\mathcal{D}}(\mathbf{z}) \exp\left\{-\beta(A(\mathbf{z}) - \hat{A}(\mathbf{z};\boldsymbol{\theta}))\right\}. \quad (40)$$

This coincides with the expression in Equation (5) and therefore the ensuing derivations hold identically. From a practical point of view, sampling need only be performed in the $\mathbf{q}$ space and therefore the adaptive SMC schemes are employed to obtain particulate approximations of the density in Equation (39). The only difference appears in the MCMC-based Rejuvenation step where the MALA sampler is employed only with regards to $\mathbf{q}$. In particular the update of Equation (32) now becomes:

$$\begin{aligned} \boldsymbol{q}_s^{(i)} - \boldsymbol{q}_{s-1}^{(i)} &= \tfrac{\Delta t_q}{2}\nabla_{\mathbf{q}}\pi_{\gamma_s}(\boldsymbol{q}_{s-1}^{(i)}) + \sqrt{\Delta t_q}\boldsymbol{r}_q \\ &= -\tfrac{\beta\Delta t_q}{2}\left(\nabla_{\mathbf{q}}V(\boldsymbol{q}_{s-1}^{(i)}) - \tfrac{\partial\hat{A}}{\partial\mathbf{z}}\nabla_{\mathbf{q}}\boldsymbol{\xi}(\mathbf{q})\right) + \sqrt{\Delta t_q}\boldsymbol{r}_q\,. \end{aligned} \quad (41)$$

It is noted that, in contrast to some ABF methods which require second-order derivatives of $\boldsymbol{\xi}$ [29], the proposed technique only needs first-order derivatives. Finally, we point out that the ability of the proposed approach to provide efficiently estimates of parametrized free energy surfaces (as in section 3.2 with respect to the temperature β), can also be exploited in the reaction coordinate case by defining a joint density:

$$p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}) \propto \exp\left\{-\beta\left(V(\mathbf{q}) + \frac{\mu}{2} \parallel \mathbf{z} - \boldsymbol{\xi}(\mathbf{q}) \parallel^2 - \hat{A}_{\mu}(\mathbf{z};\boldsymbol{\theta})\right)\right\}, \quad (42)$$

where as in [51] an artificial spring with stiffness μ has been added. Clearly for $\mu \to \infty$ one recovers the aforementioned description, but for all other values of μ the formulation reduces to that of Equation (3) where in place of $V(\mathbf{q}, \boldsymbol{z})$ we now have $V(\mathbf{q}) + \frac{\mu}{2} \parallel \mathbf{z} - \boldsymbol{\xi}(\mathbf{q}) \parallel^2$. One can therefore obtain free energy surfaces for various μ values. For smaller μ the free energy would be flatter

and in the extreme case of $\mu = 0$ it would be constant. As μ increases, the complexities of the free energy surface would become pronounced. Hence by exploiting the idea of section 3.2, a sequence of problems parametrized by μ rather than β, can be constructed to gradually move to larger μ values by using the free energy of the previous μ as an initial guess for the new one. The adaptive SMC scheme would ensure a smooth enough transition while retaining a good level of accuracy for the approximations obtained.

3.2 Obtaining the free energy landscape for various temperatures.

The methodology described in the previous sections is suitable for calculating the free energy as a function of $\mathbf{z}$ at a given temperature. However, one is often interested in the temperature dependence of the free energy landscape. In order to achieve this goal we make use of the following two facts. Firstly, the free energy landscape at higher temperatures is flatter and secondly that nearby temperatures have similar free energy landscapes. Based on these, we propose a natural extension to the sequential sampling framework of subsection 2.4 that can efficiently produce estimates of the free energy at various temperatures. The idea is to start from a higher temperature, compute the free energy as described before, then gradually move towards lower temperatures using the free energy of the previous temperature as an initial guess for the new one. In particular given the free energy estimate $\hat{A}_{\beta_1}(\boldsymbol{z}; \boldsymbol{\theta}(\beta_1))$ and the particulate approximation of the joint density $p_{\beta_1}(\mathbf{q}, \mathbf{z}|\boldsymbol{\theta}(\beta_1))$ at a temperature $1/\beta_1$, we propose employing the aforementioned adaptive SMC in order to obtain a particulate approximation of the following joint density at $\beta_2 > \beta_1$ (i.e. for

lower temperature)

$$p_{\beta_2}(\mathbf{q},\mathbf{z}\mid\boldsymbol{\theta}(\beta_1))\propto\exp\left\{-\beta_2\left(V(\mathbf{q},\mathbf{z})-\hat{A}_{\beta_1}(\boldsymbol{z};\boldsymbol{\theta}(\beta_1))\right)\right\}. \tag{43}$$

The iterations enumerated in Algorithm 2 can then be carried out in the same fashion by updating the existing $\boldsymbol{\theta}$ as well as adding new kernels if the convergence criteria are not satisfied.

The critical step involves building a sequence of distributions from $p_{\beta_1}(\mathbf{q},\mathbf{z}|\boldsymbol{\theta}(\beta_1))$ to $p_{\beta_2}(\mathbf{q},\mathbf{z}\mid\boldsymbol{\theta}(\beta_1))$ in Equation (43). For this purpose and similarly to a simulated annealing schedule we employ

$$\pi_\gamma(\mathbf{q},\mathbf{z})\propto\exp\left\{-((1-\gamma)\beta_1+\gamma\beta_2)\left(V(\mathbf{q},\mathbf{z})-\hat{A}_{\beta_1}(\boldsymbol{z};\boldsymbol{\theta}(\beta_1))\right)\right\}. \tag{44}$$

The steps in Algorithm 1 should be adjusted to the aforementioned sequence of bridging distributions with the most striking difference in the *Reweighing* step where the updated weights in Equation (31) should now be given by

$$\begin{aligned} w_s^{(i)}(\gamma_s) &= w_{s-1}^{(i)}\ \frac{\pi_{\gamma_s}(\mathbf{q}_{s-1}^{(i)},\boldsymbol{z}_{s-1}^{(i)})}{\pi_{\gamma_{s-1}}(\boldsymbol{q}_{s-1}^{(i)},\boldsymbol{z}_{s-1}^{(i)})} \\ &= w_{s-1}^{(i)}\ \exp\left\{-(\gamma_s-\gamma_{s-1})(\beta_2-\beta_1)\left(V(\mathbf{q}_{s-1}^{(i)},\boldsymbol{z}_{s-1}^{(i)})-\hat{A}_{\beta_1}(\boldsymbol{z};\boldsymbol{\theta}(\beta_1))\right)\right\}. \end{aligned} \tag{45}$$

We demonstrate the efficacy of such an approach in the last example of section 4. It is finally noted that at the beginning of iterations at each new temperature, kernels with very small weights θ_j were removed if $\frac{|\theta_j|}{\max_i|\theta_i|}\leq 0.01$.

4 Numerical Examples

In the ensuing numerical examples the following parameter values were used:

(1) SMC: $\mathrm{ESS}_{\min}=N/2,\zeta=0.95$.

(2) Time-step adaptation: We used the multiplication factors $a_i = 1.2$ to increase, and $a_d = 0.7$ to decrease the time-step so that the acceptance ratio remained between 50% and 80%.
(3) Stochastic gradient descent: $m_0 = 0$.

The rest of the parameter values are discussed in each particular problem.

The algorithm we used for the solution of the kernel addition optimization problem defined in Equation (26) was the simplex method for multidimensional minimization [60]. The centers $\mathbf{z}_j$ of the kernels were restricted within the domain $\mathcal{D}$, while the bandwidths $\tau_{j,\ell}$ were allowed to have values within $0.01\,\mathrm{diam}(\mathcal{D})$ and $0.5\,\mathrm{diam}(\mathcal{D})$ [7].

4.1 Two-Dimensional Toy Example

Consider a two-dimensional system [75,42] with a single parameter z, interacting with potential energy:

$$V(q;z) = \cos(2\pi z)(1 + d_1 q) + d_2 q^2.$$

Assume that q given z and β is distributed according to:

$$p(q|z,\beta) \propto \exp\left(-\beta V(q;z)\right),$$

where β is also a fixed parameter that plays the role of an inverse temperature. We wish to calculate an approximation $\hat{A}(z)$ of the free energy $A(z)$ on an interval $\mathcal{D} = [-0.5, 0.5]$. The true free energy can be found analytically to be

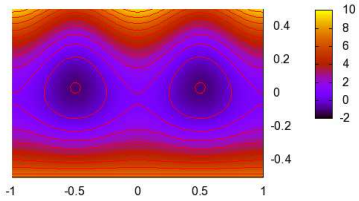
$$A(z) = \cos(2\pi z) - \frac{d_1^2 \cos(2\pi z)^2}{4d_2} + c,$$

[7] $\mathrm{diam}(\mathcal{D}) = \sup\{|\mathbf{z}_1 - \mathbf{z}_2| : \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{D}\}$ denotes the diameter of the set $\mathcal{D}$.

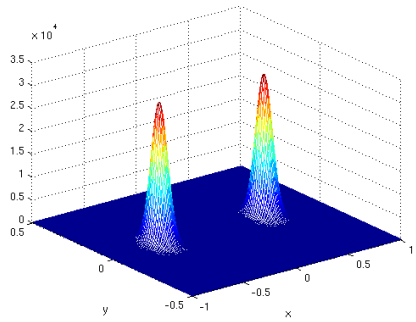
where c is a constant that depends upon the specific choice of the fixed parameters. In what follows, we choose c so that $A(-0.5) = 0$.

To demonstrate our method in this simple example we used $d_1 = 2, d_2 = 30$. The potential energy $V(q; z)$ for this choice of the parameters is depicted in Figure 1(a). We fix the inverse temperature to $\beta = 10$. As shown in Figure 1(b), the distribution is bimodal with a big region of practically zero probability separating the two modes. Hence, metastability along the parameter z is apparent. The performance of the proposed method with respect to the number of replicas used in the adaptive SMC scheme is depicted in Figures 2 and 3 which show the evolution of the estimated free energy landscape with $N = 100$ and $N = 10,000$ replicas respectively. In both cases the method is capable of capturing correctly the characteristics of the reference solution and as expected the variance of the computated solution is less when the number of replicas is larger. In both cases the Robbins-Monro learning series is picked to be $\eta_m = \eta m^{-a}$ with $a = 0.6$ and the learning rate $\eta = 0.1$.

Figure 4(a) shows the first three kernels selected by the greedy scheme described in section 2.3. Figure 4(b) depicts the log-values of the kernel weights $\{\theta_j\}_{j=1}^k$ which clearly demonstrate the ability of the proposed approach to provide sparse approximations. The first kernel selected has the greatest weight and hence it contains the majority of the information about the free energy curve. The rest of the kernels are progressive corrections of the estimate given by the first kernel. This conclusion is also supported by the result of Figure 5 which shows the evolution of the reduction in the KL divergence with respect to the total number of iterations as quantified by adding the Δ_{k+1} in Equation (23). Clearly the first kernel offers the greatest KL gain (Δ_1) and further kernel additions offer (almost always) progressively smaller reductions in the



(a) The potential energy $V(q_1, q_2)$ for $d_1 = 2, d_2 = 30$



(b) The probability distribution $p(q_1, q_2|\beta)$ for $d_1 = 2, d_2 = 30, \beta = 10$

Fig. 1. Potential energy and pdf for the toy example of section 4.1

KL divergence.

4.2 WCA Dimer

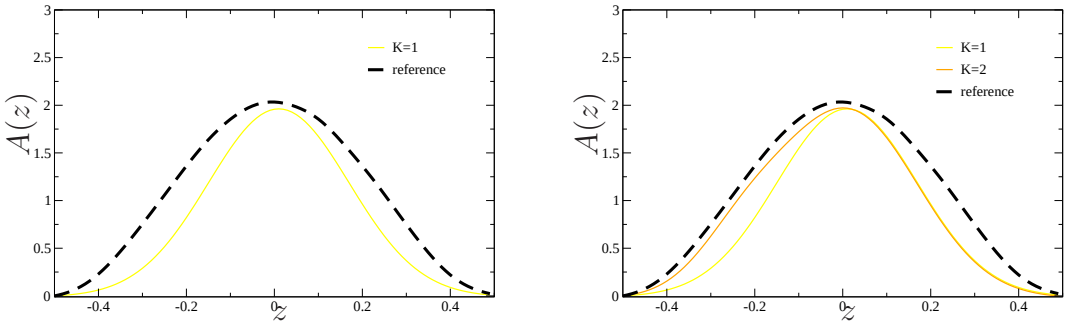
We consider $n = 16$ atoms in a two-dimensional fully periodic box of side l which interact with a purely repulsive WCA pair potential [42]:

$$V_{\mathrm{WCA}}(r) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \epsilon & \text{, if } r \geq r_0 \\ 0 & \text{, otherwise} \end{cases},$$

where $r_0 = 2^{1/6}\sigma$. The parameters σ and ϵ give the length and energy scales respectively. Two of these atoms (say atoms 1 and 2) are assumed to form a dimer and their interaction is described instead with a double-well potential:

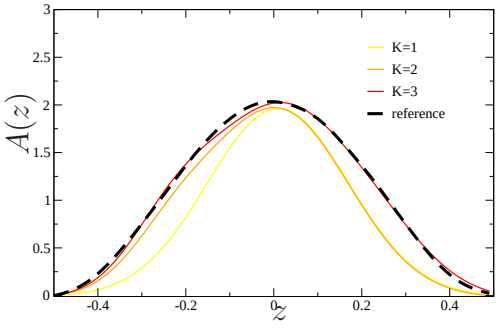
$$V_S(r) = h\left[1 - \frac{(r - r_0 - w)^2}{w^2}\right],$$

where h, w are fixed parameters and r the distance between them. This potential has two equilibrium points r_0 and $r_0 + 2w$. The barrier separating the two equilibria is h.

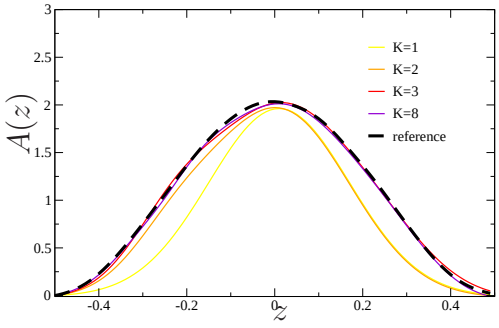


(a) k=1

(b) k=2



(c) k=3



(d) k=8

Fig. 2. Free energy profiles for various kernel numbers k when using $N = 100$ replicas in the adaptive SMC scheme

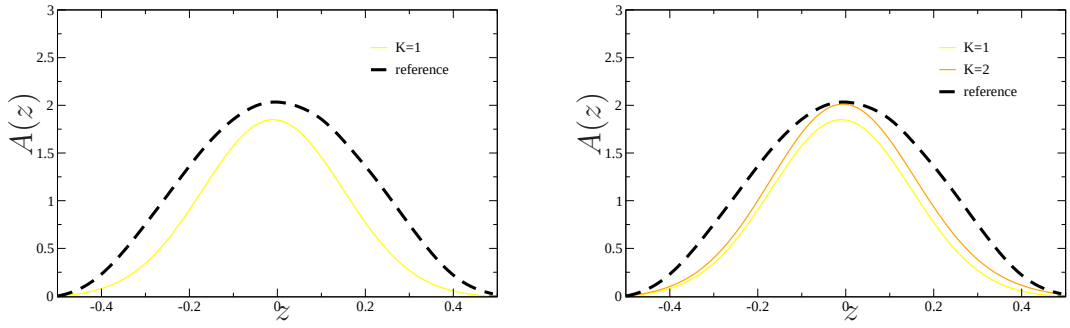
Let $\boldsymbol{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N)$ with $\boldsymbol{q}_i \in \mathbb{R}^2$ denoting the position of atom i. The potential energy of the system is

$$V(\boldsymbol{q}) = V_S(|\boldsymbol{q}_1 - \boldsymbol{q}_2|) + \sum_{i=1}^{2}\sum_{j=3}^{N} V_{\mathrm{WCA}}(|\boldsymbol{q}_i - \boldsymbol{q}_j|) + \sum_{2<i<j} V_{\mathrm{WCA}}(|\boldsymbol{q}_i - \boldsymbol{q}_j|).$$

We consider an NVT ensemble (the volume V is determined by the side of the box l). The probability distribution of the atomic positions $\boldsymbol{q}$ is:

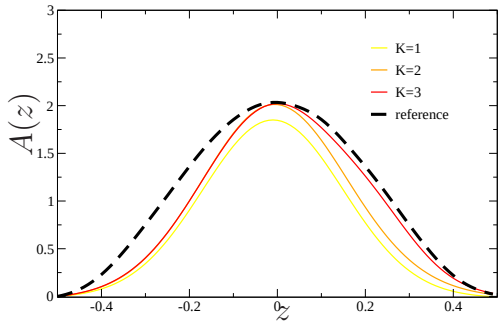
$$p(\boldsymbol{q}|\beta) \propto \exp\left(-\beta V(\boldsymbol{q})\right),$$

where $\beta = \frac{1}{k_B T}$, k_B is the Boltzmann constant and T the temperature of the system. Under these assumptions atoms 1 and 2 will form a dimer with two equilibrium lengths. An *effective potential* of the dimer length in the presence

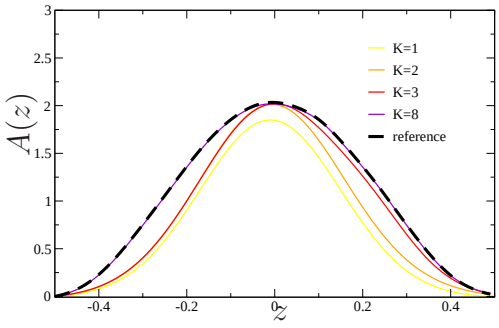


(a) k=1

(b) k=2



(c) k=3



(d) k=8

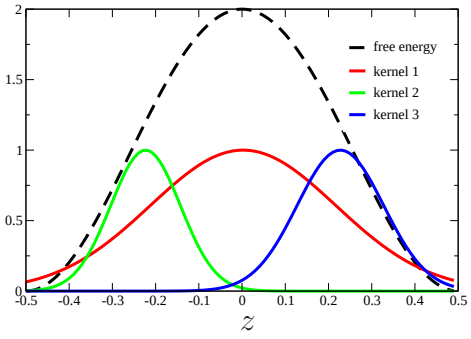
Fig. 3. Evolution of free energy estimates at various kernel numbers k when using $N = 10,000$ replicas in the adaptive SMC scheme

of the other atoms is given by the free energy $A(r)$ with respect to the reaction coordinate:

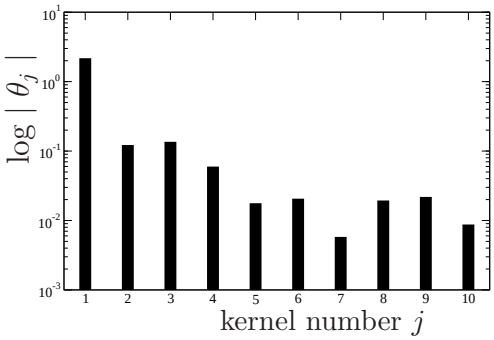
$$z = \xi(\boldsymbol{q}) = \| \boldsymbol{q}_1 - \boldsymbol{q}_2 \|_2,$$

where $|| \cdot ||_2$ is the Euclidean norm of $\mathbb{R}^2$.

We calculate $A(z)$ using our scheme for two different box sizes (densities): $l = 5$ (high density) and $l = 12$ (low density). The parameters are set to $n = 16$ atoms, $\beta = 1, \epsilon = 1, \sigma = 1, h = 1, w = 0.5$. We employed $N = 500$ replicas and the Robbins-Monroe learning series is again $\eta_m = \eta m^{-a}$ with $a = 0.501$ and $\eta = 0.1$. The resulting free energy curves at various stages of the estimation process with increasing number of kernels are shown in Figure 6. Figure 7



(a) First three kernels $K_j(.)$ picked by the greedy optimization scheme to illustrate selection of location and bandwidth parameters ($N = 10,000$).



(b) Log absolute weights ($\log \mid \theta_j \mid$) of the first 10 kernels added. The θ value of the first kernel is over one order of magnitude larger than the rest.

Fig. 4. Kernels selected and kernel weights obtained with $N = 10,000$ replicas

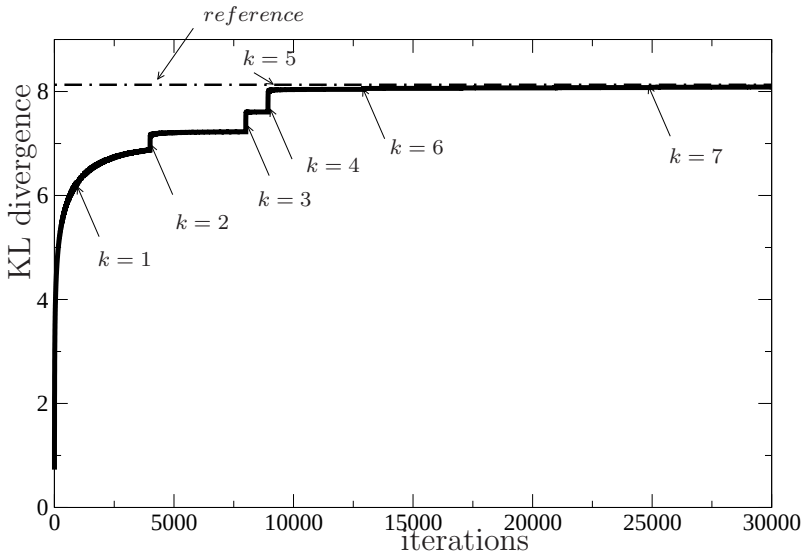
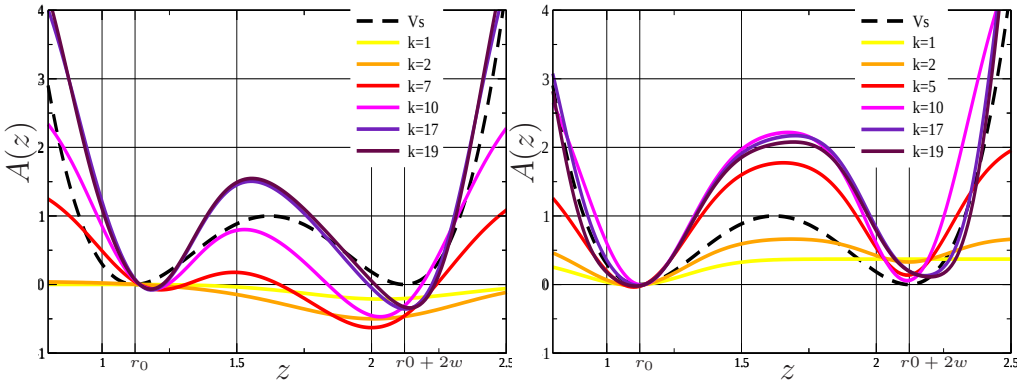


Fig. 5. Evolution of the KL divergence gain Δ_{k+1} (Equation (23)) at various k with respect to the total number of gradient ascent iterations performed. The dashed line (reference) depicts the KL-distance between the initial distribution ($\boldsymbol{\theta} = \mathbf{0}$) and the target uniform distribution which can be calculated analytically for this example.



(a) Low Density $l = 12$. (b) High Density $l = 5$.

Fig. 6. The free energy of the dimer at two different densities and for various numbers of kernels k, compared with $V_S(r)$. Notice that at low density (a) the right well becomes the most probable. This situation is reversed at high density (b).

compares the empirical cumulative distribution function of the replicas with that of the target uniform. Their proximity indicates convergence as it can also be established by a Kolmogorov-Smirnov test.

From a physical point of view, we notice that at *low density* i.e. when the box size is $l = 12$ (Figure 6(a)), the equilibria move to the right with the well closest to $r_0 + 2w$ becomes the most probable. Furthermore the free energy barrier is slightly decreased as compared to the *high density* case when $l = 5$ (Figure 6(b)). Under these conditions the equilibria move to the left and the well closest to r_0 becomes the most probable.

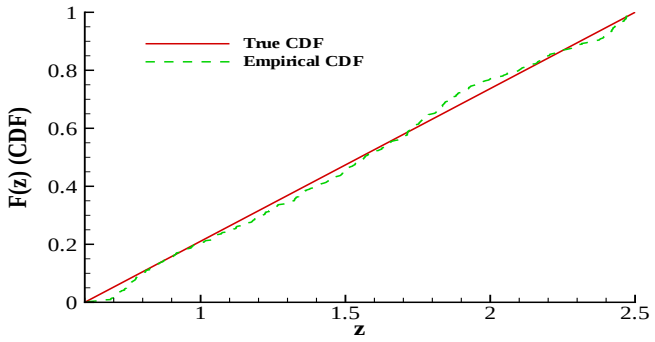


Fig. 7. Comparison of the empirical cumulative distribution function of the reaction coordinates of the replicas with that of the target uniform for $l = 5$. We also performed a Kolmogorov-Smirnov test [54] as implemented in Matlab [55]. The default confidence level was $h = 0.05$. The hypothesis was accepted and the test reported an asymptotic p value of 0.3515. The value of the test statistic was 0.0631.

4.3 38-Atom Lennard-Jones Cluster (LJ_{38})

We consider a 38-atom cluster in 3-dimensional space with pairwise interactions given by the Lennard-Jones potential:

$$V_{\mathrm{LJ}}(r) = \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right], \tag{46}$$

with ϵ and σ playing the role of energy and length scale respectively. Let the Cartesian coordinates of the system be:

$$\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{38}), \mathbf{q}_i \in \mathbb{R}^3. \tag{47}$$

Then the potential energy of the system is:

$$V(\mathbf{q}) = \sum_{i<j} V_{\mathrm{LJ}}\left(|\mathbf{q}_i - \mathbf{q}_j|\right).$$

Finally we assume that the replicas follow an NVT distribution of the form:

$$p(\mathbf{q}|\beta) \propto \exp\left\{-\beta V(\mathbf{q})\right\},$$

(a) $Q_4 \approx 0.01$. (b) $Q_4 \approx 0.19$ (truncated octahedron).

Fig. 8. Indicative metastable states corresponding to the two wells of the free energy landscape with respect to order parameter Q_4 (Equation (48)).

where $\beta = 1/k_B T$. At zero temperature the system is known to have a global minimum yielding an FCC truncated octahedron (Figure 8(b)). The second and third lower energies give incomplete Mackey icosahedra. Furthermore there is a big number of liquid-like local minima ([25,7]).

Consider the family of order parameters initially introduced in [69]:

$$Q_l = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{Q}_{lm}|^2 \right)^2, \tag{48}$$

with:

$$\bar{Q}_{lm} = \frac{1}{N_b} \sum_{r_{ij}<r_0} Y_{lm}(\theta_{ij}, \phi_{ij}),$$

where the sum is over all the N_b pairs of atoms with $r_{ij} = |\mathbf{q}_i - \mathbf{q}_j| < r_0$, $Y_{lm}(\theta, \phi)$ is a spherical harmonic, while θ_{ij} and ϕ_{ij} are the polar and azimuthal angles of a bond vector with respect to an arbitrary coordinate system. In [7] it is shown that for $l = 5$, Q_4 can distinguish the FCC structure but not the icosahedral and liquid-like minima (Figure 8(a)). However, if one also

considers the potential energy as a reaction coordinate, the two structures are well-separated. Hence, we define the two dimensional reaction coordinate:

$$\xi(\mathbf{q}) = (Q_4(\mathbf{q}), V(\mathbf{q}))\,.$$

and compute the free energy:

$$A(Q_4, E) = \beta^{-1} \int \exp\left\{-\beta V(\mathbf{q})\right\} \delta(Q_4 - Q_4(\mathbf{q}))\delta(E - V(\mathbf{q}))d\mathbf{q},$$

over the domain:

$$\mathcal{D} = [0, 0.2] \times [-175\epsilon, -145\epsilon],$$

for a temperature range $k_BT = 0.21$ to $k_BT = 0.091$ using the tempering scheme described in Section 3.2. We employ $N = 100$ replicas and 10 MCMC/Rejuvenation steps per replica. At each $\beta = k_BT$, the Robbins-Monro learning series was adjusted to $\eta_m = \eta m^{-a}$ with $a = 0.501$ and a learning rate $\eta = 0.1/\beta$. The adaptive SMC scheme automatically determined 260 intermediate steps/distributions in order to cover the whole range of the aforementioned temperatures. The time step Δt_q employed in the MALA sampler was adaptively adjusted as discussed previously and took values between 10^{-4} (low temperatures) and 7×10^{-4} (high temperatures). The very first step, at $T = 0.21$ ($\beta = 4.76$) required $12,000$ optimization iterations to converge with a cost of approximately 7.2×10^5 time steps per replica. It is emphasized that due to the parallelizable nature of the SMC scheme employed, each replica can be simulated on a different CPU. Centralized control is only required for the evaluation of the ESS (Reweighting step) and during the Resampling step which are computationally less expensive than the Rejuvenation step.

The sequence of intermediate β's determined automatically by the scheme discussed in section 3.2 is depicted in Figure 9. The similarity of the free en-

ergy surfaces at neighboring temperatures allowed us to converge with, on average, 800 optimization iterations at each intermediate β. The overall simulation cost amount to 2.4×10^4 time steps per replica. This was approximately 30 times less than the 7.0×10^5 time steps per replica which were needed when calculating the free energy solely at the final $\beta = 0.091$ i.e. without using the sequence of temperatures and corresponding free energies. The significant reduction indicates the potential computational savings afforded by the sequential approach advocated in this work.

The free energy surfaces computed are depicted in Figure 10 at four indicative temperatures. The number of kernels selected by the algorithm varied between 90 and 120. As it has been reported in previous studies [7], we identified two metastable states at $Q_4 \approx 0.01$ which corresponds to the truncated octahedron and at $Q_4 \approx 0.19$ which corresponds to the icosahedron. The latter becomes more pronounced at lower temperatures.

In order to assess the quality of the results in two dimensions we also calculated the free energy profile using only Q_4 as the reaction coordinate (Figure 11) and compared it with the result obtained by performing a numerical integration of the two-dimensional free energy surface i.e. by computing $A(Q_4) = -\beta^{-1} \log \int e^{-\beta A(Q_4,E)} dE$. The two free-energy curves are depicted in Figure 11 where good agreement is observed at two different temperatures.

5 Conclusions

In summary, the proposed method provides a unifying framework for estimating the free energy function simultaneously with biasing the dynamics. The

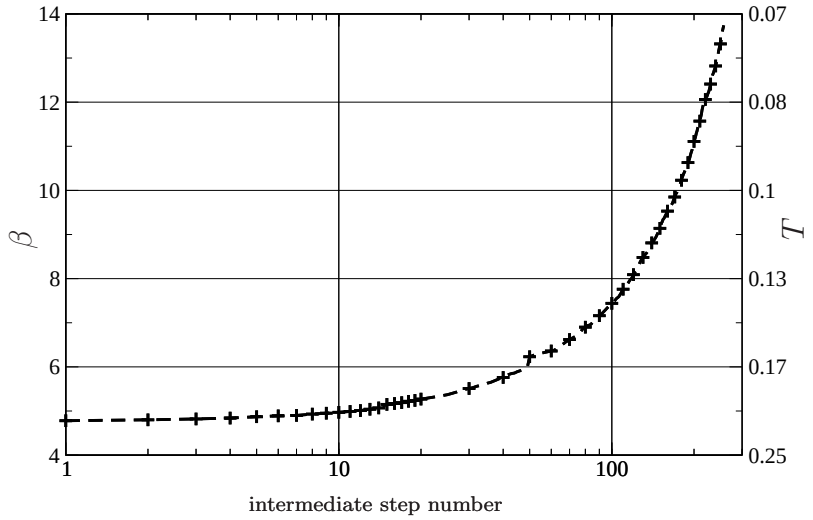
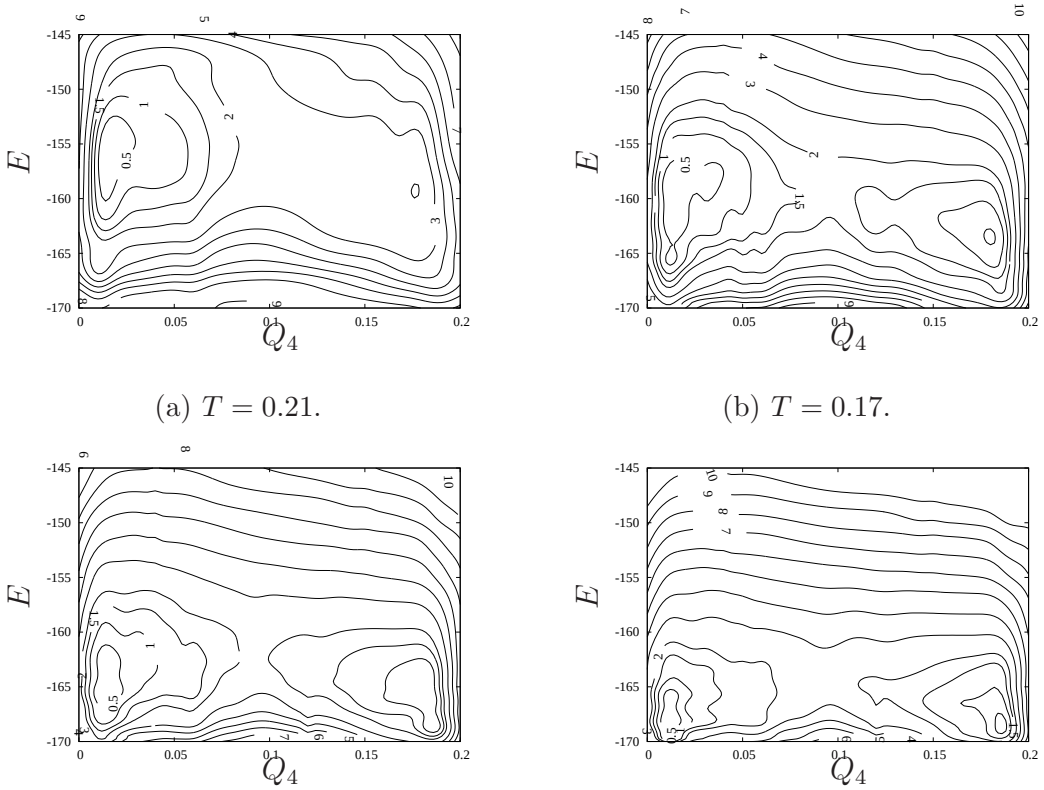


Fig. 9. Sequence of intermediate β's identified by the scheme discussed in section 3.2 for the LJ_{38} cluster. The free energy landscape was calculated at each of these temperatures by efficiently updating the free energy surface at the previous step.

minimization of the Kullback-Leibler divergence in the extended space provides rigorous convergence bounds and diagnostics. It requires minimal adjustment of parameter values a priori (basically only the learning rate λ and convergence tolerances) as it is adaptive and automatically promotes sparse representations of the free energy surface. The proposed approach shares a common theme with other adaptive methods in free-energy estimation and Monte Carlo methods in general, in that the target distribution (in our case $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$) is modified from iteration to iteration based on its past samples [48]. The approximation of the free energy $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$, biases the potential of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ (Equation (3)) and allows the system to overcome free energy barriers [53]. As in [24], no binning is needed and the bias potential is nonlocal, providing information about the free energy landscape not only at the states visited but in their neighborhood as well. It offers several possibilities for further improvements by considering different optimization schemes (e.g. noisy



(a) $T = 0.21$. (b) $T = 0.17$.

(c) $T = 0.14$. (d) $T = 0.11$.

Fig. 10. Free energy contours $A(Q_4, E)$ with respect to the two reaction coordinates Q_4 (x-axis) and E (y-axis) at various temperatures for LJ_{38}.

conjugate gradients) and employing different basis functions (e.g. wavelets). Its sequential nature allows the efficient computation of a family of free energy surfaces at different temperatures. We believe that these features make the proposed approach suitable to calculate the free energy of systems more physically challenging than the ones discussed in this paper.

Acknowledgement We are much indebted to Dr. Florent Calvo for providing us with the Q4-gradient subroutine for the third numerical example. This work was supported by the OSD/AFOSR MURI'09 award to Cornell University on uncertainty quantification

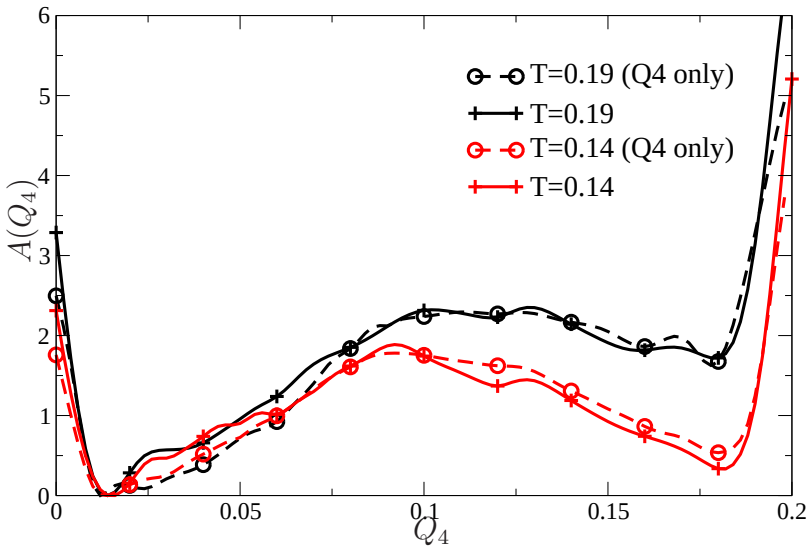


Fig. 11. Free energy profiles $A(Q_4)$ obtained using the proposed scheme while using only reaction coordinate (dashed lines) and by integrating the two dimensional free energy surface i.e. $A(Q_4) = -\beta^{-1} \log \int e^{-\beta A(Q_4,E)} dE$ (solid lines) for two temperatures $T = 0.19$ and $T = 0.14$.

Appendix A

This appendix contains a proof of the second sufficient condition C2 for the almost sure convergence of the proposed stochastic approximation scheme as discussed in Section 2.2.

For a fixed number of kernels to k, let $\tilde{\boldsymbol{\theta}}^k \in \boldsymbol{\Theta}$ be the unique global minimum of $I(\boldsymbol{\theta}^k)$, let $0 < \epsilon < 1$ and define the subset of $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta}_\epsilon = \left\{ \boldsymbol{\theta}^k \in \boldsymbol{\Theta} : \epsilon < \| \boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k \| < 1/\epsilon \right\}.$$

Our goal is to prove that:

$$\inf_{\boldsymbol{\theta}^k \in \boldsymbol{\Theta}_\epsilon} \left(\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k \right)^T \boldsymbol{J}(\boldsymbol{\theta}^k) = \inf_{\epsilon < \|\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k\| < 1/\epsilon} \left(\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k \right)^T \boldsymbol{J}(\boldsymbol{\theta}^k) > 0. \tag{49}$$

To this end, observe that:

$$\begin{aligned} Z(\tilde{\boldsymbol{\theta}}^k) &= \textstyle\int_{\mathcal{D}\times\mathcal{M}} e^{-\beta\left(V(\mathbf{q};\mathbf{z}) - \hat{A}(\mathbf{z};\tilde{\boldsymbol{\theta}}^k)\right)}\, d\mathbf{q}d\mathbf{z} \\ &= Z(\boldsymbol{\theta}^k) E_{p(\mathbf{z}|\boldsymbol{\theta}^k)} \left[e^{\beta\left(\hat{A}(\mathbf{z};\tilde{\boldsymbol{\theta}}^k) - \hat{A}(\mathbf{z};\boldsymbol{\theta}^k)\right)} \right] \end{aligned} \tag{50}$$

and as a result of Jensen's inequality:

$$\begin{aligned} \log \tfrac{Z(\tilde{\boldsymbol{\theta}}^k)}{Z(\boldsymbol{\theta}^k)} &= \log E_{p(\mathbf{z}|\boldsymbol{\theta}^k)} \left[e^{\beta\left(\hat{A}(\mathbf{z};\tilde{\boldsymbol{\theta}}^k) - \hat{A}(\mathbf{z};\boldsymbol{\theta})\right)} \right] \\ &\geq \beta E_{p(\mathbf{z}|\boldsymbol{\theta}^k)} \left[\hat{A}(\mathbf{z};\tilde{\boldsymbol{\theta}}^k) - \hat{A}(\mathbf{z};\boldsymbol{\theta}^k) \right] \end{aligned} \tag{51}$$

Hence, the difference between the values of the objective function at the op-

timum point $\tilde{\boldsymbol{\theta}}^k$ and an arbitrary $\boldsymbol{\theta}^k \in \boldsymbol{\Theta}_\epsilon$ is non-negative and:

$$\begin{aligned}
0 \leq I(\boldsymbol{\theta}^k) - I(\tilde{\boldsymbol{\theta}}^k) &= -\int_{\mathcal{D}} \pi(\mathbf{z}) \log p(\mathbf{z}|\boldsymbol{\theta}^k)\, d\mathbf{z} + \int_{\mathcal{D}} \pi(\mathbf{z}) \log p(\mathbf{z}|\tilde{\boldsymbol{\theta}}^k)\, d\mathbf{z} \\
&= \beta E_{\pi(\mathbf{z})}\left[\hat{A}(\mathbf{z};\tilde{\boldsymbol{\theta}}^k) - \hat{A}(\mathbf{z};\boldsymbol{\theta}^k)\right]\, d\mathbf{z} - \log \tfrac{Z(\tilde{\boldsymbol{\theta}}^k)}{Z(\boldsymbol{\theta}^k)} \\
&\leq \beta\left(E_{\pi(\mathbf{z})}\left[\hat{A}(\mathbf{z};\tilde{\boldsymbol{\theta}}^k) - \hat{A}(\mathbf{z};\boldsymbol{\theta}^k)\right] - E_{p(\mathbf{z}|\boldsymbol{\theta}^k)}\left[\hat{A}(\mathbf{z};\tilde{\boldsymbol{\theta}}^k) - \hat{A}(\mathbf{z};\boldsymbol{\theta}^k)\right]\right) \\
&= -\textstyle\sum_{j=1}^k (\tilde{\theta}_j^k - \theta_j^k)\beta\left[E_{p(\mathbf{z}|\boldsymbol{\theta}^k)}[K_j(\mathbf{z})] - E_{\pi(\mathbf{z})}[K_j(\mathbf{z})]\right] \\
&= (\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k)^T \boldsymbol{J}(\boldsymbol{\theta}^k)
\end{aligned} \tag{52}$$

where the last equality is a result of the definition of the gradient $\boldsymbol{J}(\boldsymbol{\theta})$ of $I(\boldsymbol{\theta})$ in Equation (15). Given that $(\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}^k)^T \boldsymbol{J}(\boldsymbol{\theta}^k) \geq I(\boldsymbol{\theta}^k) - I(\tilde{\boldsymbol{\theta}}^k)$, it suffices to prove that:

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon} \left(I(\boldsymbol{\theta}^k) - I(\tilde{\boldsymbol{\theta}}^k)\right) > 0. \tag{53}$$

We do this by contradiction. Suppose that:

$$\inf_{\boldsymbol{\theta}^k \in \boldsymbol{\Theta}_\epsilon} \left(I(\boldsymbol{\theta}^k) - I(\tilde{\boldsymbol{\theta}}^k)\right) = 0 \tag{54}$$

Then, there exists a sequence $\{\boldsymbol{\theta}_m^k\} \subset \boldsymbol{\Theta}_\epsilon$ such that:

$$I(\boldsymbol{\theta}_m^k) \to I(\tilde{\boldsymbol{\theta}}^k), \text{ as } m \to \infty. \tag{55}$$

Since the sequence $\{\boldsymbol{\theta}_m^k\}$ is bounded, it has a convergent subsequence $\{\boldsymbol{\theta}_{m_n}^k\}$, i.e

$$\boldsymbol{\theta}_{m_n}^k \to \boldsymbol{\theta}^*, \text{ as } n \to \infty,$$

for some $\boldsymbol{\theta}^* \in \text{clos}(\boldsymbol{\Theta}_\epsilon)$ [8]. From the continuity of $I(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we must also have:

$$I(\boldsymbol{\theta}_{m_n}^k) \to I(\boldsymbol{\theta}^*)$$

[8] $\text{clos}(A)$ denotes the closure of the set $A \subset \mathbb{R}^k$ under the Eucledian metric.

From the uniqueness of the limit in Equation (55), we get that:

$$I(\tilde{\boldsymbol{\theta}}^k) = I(\boldsymbol{\theta}^*)$$

Since $I(\boldsymbol{\theta})$ has a unique global minimum:

$$\tilde{\boldsymbol{\theta}}^k = \boldsymbol{\theta}^* \in \text{clos}(\boldsymbol{\Theta}_\epsilon),$$

which contradicts Equation (54).

Appendix B

This appendix contains an alternative interpretation of the proposed formulation (section 2.1) through the prism of the well-known Expectation-Maximization scheme (EM, [23]) and offers some potential Bayesian extensions.

For that purpose we define the function $I'(\boldsymbol{\theta}) = -I(\boldsymbol{\theta}) = \int_{\mathcal{D}} \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta}) d\mathbf{z}$ (Equation (12)). Maximizing $I'(\boldsymbol{\theta})$ is equivalent to minimizing $I(\boldsymbol{\theta})$. We note that for all $\mathbf{z}$ and any density $Q(\mathbf{q}),\ \mathbf{q} \in \mathcal{M}$:

$$\begin{aligned}
\log p(\boldsymbol{z} \mid \boldsymbol{\theta}) &= \log \textstyle\int_{\mathcal{M}} p(\boldsymbol{q}, \boldsymbol{z} \mid \boldsymbol{\theta})\ d\boldsymbol{q} \\
&= \log \textstyle\int_{\mathcal{M}} Q(\boldsymbol{q}) \frac{p(\boldsymbol{q}, \boldsymbol{z} \mid \boldsymbol{\theta})}{Q(\boldsymbol{q})}\ d\boldsymbol{z} \\
&\geq \textstyle\int_{\mathcal{M}} Q(\boldsymbol{q}) \log \frac{p(\boldsymbol{q}, \boldsymbol{z} \mid \boldsymbol{\theta})}{Q(\boldsymbol{q})}\ d\boldsymbol{z} \quad (Jensen's\ inequality) \\
&= \textstyle\int_{\mathcal{M}} Q(\boldsymbol{q}) \log p(\boldsymbol{q}, \boldsymbol{z} \mid \boldsymbol{\theta})\ d\boldsymbol{z} - \int Q(\boldsymbol{q}) \log Q(\boldsymbol{q})\ d\boldsymbol{q} \\
&= \mathcal{F}(Q, \boldsymbol{\theta}; \boldsymbol{z})
\end{aligned} \tag{56}$$

This lower bound $\mathcal{F}$ holds for all $\mathbf{z} \in \mathcal{D}$ and depends on the auxiliary density $Q(\mathbf{q})$ and the model parameters $\boldsymbol{\theta}$. Furthermore it also provides a lower bound on $I'(\boldsymbol{\theta})$:

$$I'(\boldsymbol{\theta}) \geq \int_{\mathcal{D}} \pi(\boldsymbol{z}) \left(\mathcal{F}(Q, \boldsymbol{\theta}; \boldsymbol{z})\right)\ d\boldsymbol{z} \tag{57}$$

For any $\boldsymbol{z} \in \mathcal{D}$ and fixed $\boldsymbol{\theta}$, the optimal density Q is:

$$Q^{opt}(\boldsymbol{q} \mid \boldsymbol{\theta}) = p(\boldsymbol{q} \mid \boldsymbol{z}, \boldsymbol{\theta}) \tag{58}$$

in which case the inequalities in Equations (56) and (57) become equalities.

More importantly though the introduction of the auxiliary density Q suggests an iterative algorithm for finding the optimal parameter values for $\boldsymbol{\theta}$. Initializing with an value $\boldsymbol{\theta}_0$, at each subsequent iteration m we alternate between the following two steps:

- Expectation step (E-step):

$$Q_m(\boldsymbol{z}) = \arg\max_{Q} \mathcal{F}(Q, \boldsymbol{\theta}_m; \boldsymbol{z}) = Q^{opt}(\boldsymbol{q} \mid \boldsymbol{\theta}_{m-1}) = p(\boldsymbol{q} \mid \boldsymbol{z}, \boldsymbol{\theta}_{m-1}), \quad \forall \boldsymbol{z} \in \mathcal{D} \tag{59}$$

- Maximization step (M-step):

$$\begin{aligned} \boldsymbol{\theta}_m &= \arg\max_{\boldsymbol{\theta}} \textstyle\int_{\mathcal{D}} \pi(\boldsymbol{z}) \left(\mathcal{F}(Q_m, \boldsymbol{\theta}; \boldsymbol{z})\right) \, d\boldsymbol{z} \\ &= \arg\max_{\boldsymbol{\theta}} \textstyle\int_{\mathcal{D}} \left(\int_{\mathcal{M}} p(\boldsymbol{q} \mid \boldsymbol{z}, \boldsymbol{\theta}_{m-1}) \log p(\boldsymbol{q}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, d\mathbf{q}\right) \pi(\boldsymbol{z}) \, d\boldsymbol{z} \\ &= \arg\max_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \boldsymbol{\theta}_{m-1}) = \arg\max_{\boldsymbol{\theta}} I'(\boldsymbol{\theta}) \end{aligned} \tag{60}$$

This essentially suggests a coordinate ascent that alternates between $\boldsymbol{\theta}$ and $Q(\boldsymbol{q})$ which is guaranteed to converge to a local maximum [57,59]. We show in the sequel that the first and second order derivatives of $H(\boldsymbol{\theta}, \boldsymbol{\theta}_{m-1})$ coincide with those of $I'(\boldsymbol{\theta})$. As a result they have a unique and identical maximum.

In particular, substituting from Equation (3), we obtain:

$$\begin{aligned} H(\boldsymbol{\theta}, \boldsymbol{\theta}_{m-1}) &= \textstyle\int_{\mathcal{D}} \left(\int_{\mathcal{M}} p(\boldsymbol{q} \mid \boldsymbol{z}, \boldsymbol{\theta}_{m-1}) (-\beta \left(V(\mathbf{q}, \boldsymbol{z}) - \hat{A}(\boldsymbol{z}; \boldsymbol{\theta})\right) - \log Z(\boldsymbol{\theta})\right) \, \pi(\boldsymbol{z}) d\boldsymbol{z} \\ &= \textstyle\int_{\mathcal{D}} \left(E_{p(\boldsymbol{q}|\boldsymbol{z}, \boldsymbol{\theta}_{m-1})} \left[-\beta \left(V(\mathbf{q}, \boldsymbol{z}) - \hat{A}(\boldsymbol{z}; \boldsymbol{\theta})\right)\right]\right) \, \pi(\boldsymbol{z}) d\boldsymbol{z} - \log Z(\boldsymbol{\theta}) \end{aligned} \tag{61}$$

Using Equation (14) and the expansion of Equation (7) for $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$, we obtain

the gradient of $\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \boldsymbol{\theta}_{m-1})$ with respect to $\boldsymbol{\theta}$:

$$\begin{aligned}
\frac{\partial H(\boldsymbol{\theta},\boldsymbol{\theta}_{m-1})}{\partial \theta_j} &= \int_{\mathcal{D}} \beta \frac{\partial \hat{A}(\boldsymbol{z};\boldsymbol{\theta})}{\partial \theta_j}\ \pi(\boldsymbol{z}) d\boldsymbol{z} - \frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_j} \\
&= \beta E_{\pi(\mathbf{z})}\left[\frac{\partial \hat{A}(\boldsymbol{z};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] - \beta - \frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_j} \\
&= \beta \left(E_{\pi(\mathbf{z})}\left[K_j(\boldsymbol{z})\right] - E_{p(\mathbf{z}|\boldsymbol{\theta})}\left[K_j(\boldsymbol{z})\right]\right)
\end{aligned} \tag{62}$$

Furthermore, the Hessian of the objective function $H(\boldsymbol{\theta}, \boldsymbol{\theta}_{m-1})$ is proportional to the covariance between the kernels i.e.:

$$\begin{aligned}
\frac{\partial^2 H(\boldsymbol{\theta},\boldsymbol{\theta}_{m-1})}{\partial \theta_j \partial \theta_l} &= -\frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l} \\
&= -\beta^2 Cov_{p(\boldsymbol{z}|\boldsymbol{\theta})}[K_j, K_l]
\end{aligned} \tag{63}$$

A Bayesian extension to the aforementioned formulation could be readily obtained by the introduction of a prior density on $\boldsymbol{\theta}$, i.e. $p(\boldsymbol{\theta})$. In this case the objective function $I'(\boldsymbol{\theta}) = -I(\boldsymbol{\theta})$ (Equation (11)) can be interpreted as the limiting log-likelihood in the case of infinite "observations" $\{\mathbf{z}_i\}_{i=1}^n$ $(n \to \infty)$ from the uniform density $\pi(\mathbf{z})$ on $\mathcal{D}$, i.e.:

$$\sum_{i=1}^{n} \log p(\mathbf{z}_i \mid \boldsymbol{\theta}) \to \int_{\mathcal{D}} \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta})\ d\mathbf{z} \tag{64}$$

Hence the log-posterior (conditioned on the "observations" $\mathbf{z}_i$) would be:

$$\begin{aligned}
\hat{I}(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta} \mid \{\mathbf{z}_i\}_{i=1}^{n\to\infty}) &\propto \underbrace{\int_{\mathcal{D}} \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta})\ d\mathbf{z}}_{log-likelihood} + \underbrace{\log p(\boldsymbol{\theta})}_{log-prior} \\
&= I'(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})
\end{aligned} \tag{65}$$

A maximization of the log-posterior $\hat{I}(\boldsymbol{\theta})$ amounts to a MAP (Maximum A Posteriori) point estimate of $\boldsymbol{\theta}$. The gradient and Hessian of $\hat{I}(\boldsymbol{\theta})$ would then be

penalized/regularized versions of the respective gradient and Hessian of $I'(\boldsymbol{\theta})$. Appropriate prior modeling could provide interesting extensions in the context of sparse representations ([27,73]). Prior modeling could also be extended ed to the remaining parameters of the expansion e.g. kernel bandwidths.

References

[1] C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal of Control and Optimization*, 44:283–312, 2005.

[2] Y. F. Atchadé and J. S. Liu. The Wang-Landau algorithm in general state spaces: Applications and convergence analysis. *Statistica Sinica*, 20(1):209—-233, 2010.

[3] B. Berg and T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12, January 1992.

[4] A. Berger. The improved iterative scaling algorithm: A gentle introduction. Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1997.

[5] J. Besag. Markov chain Monte Carlo for statistical inference. Technical report, Center for Statistics and the Social Sciences, University of Washington, Seattle, WA, 2001.

[6] G. Bussi, A. Laio, and M. Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Physical Review Letters*, 96(9):090601, 2006.

[7] F. Calvo, J. P. Neirotti, David L. Freeman, and J. D. Doll. Phase changes in 38-atom Lennard-Jones clusters. II. A parallel tempering study of equilibrium and

dynamic properties in the molecular dynamics and microcanonical ensembles. *The Journal of Chemical Physics*, 112(23):10350–10357, 2000.

[8] E. Cancès, F. Legoll, and G Stoltz. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41:351–389, 2007.

[9] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[10] P. Carbonetto, M. King, and F. Hamze. A stochastic approximation method for inference in probabilistic graphical models. In *NIPS 22*, page 216224, 2009.

[11] C. Chipot and A. Pohorille. *Free Energy Calculations*. Springer, 2007.

[12] N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to bayesian inference. *Annals of Statistics*, 32(6):2385–2411, 2004.

[13] G. Ciccotti, T. Lelièvre, and E. Vanden-Eijnden. Sampling Boltzmann-Gibbs distributions restricted on a manifold with diffusions: Application to free energy calculations. *Rapport de recherche du CERMICS*, 309, 2006.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John wiley and Sons, 2nd edition, 1991.

[15] D. Crisan and A. Doucet. Convergence of sequential Monte Carlo methods. Technical report, Technical Report CUED/FINFENG/TR381, Signal Processing Group, Department of Engineering, University of Cambridge, 2000.

[16] E. Darve and A. Pohorille. Calculating free energies using average force. *The Journal of Chemical Physics*, 115(20):9169, 2001.

[17] E. Darve, D. Rodriguez-Gómez, and A. Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *Journal of Chemical Physics*, 128(14):144120, 2008.

[18] P. Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[19] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo for Bayesian computation. In *Bayesian Statistics 8*. Oxford University Press, 2006.

[20] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B*, 68(3):411–436, 2006.

[21] P. Del Moral, A. Doucet, and A. Jasra. On adaptive resampling procedures for sequential monte carlo methods. *Bernoulli (to appear)*, 47, 2010.

[22] S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[24] B. M. Dickson, F. Legoll, T. Lelièvre, G. Stoltz, and P. Fleurat-Lessard. Free energy calculations: an efficient adaptive biasing potential method. *The Journal of Physical Chemistry B*, 114(17):5823–5830, May 2010.

[25] J. P. K. Doye, M. A. Miller, and D. J. Wales. The double-funnel energy landscape of the 38-atom Lennard-Jones cluster. *The Journal of Chemical Physics*, 110(14):6896–6906, 1999.

[26] M. Fasnacht, R. Swendsen, and J. Rosenberg. Adaptive integration method for Monte Carlo simulations. *Physical Review E*, 69(5):056704, 2004.

[27] M. A. T. Figueiredo. Adaptive Sparseness Using Jeffreys Prior. In *NIPS*, 2001.

[28] A. Gelman and X. L. Meng. Simulating normalizing constants: from importance

sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, May 1998.

[29] J. Hénin and C. Chipot. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *The Journal of Chemical Physics*, (121):2904–2914, 2004.

[30] J. E. Hunter, W. P. Reinhardt, and T. F. Davis. A finite-time variational method for determining optimal paths and obtaining bounds on free energy changes from computer simulations. *The Journal of Chemical Physics*, 99(9):6856, 1993.

[31] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56(5):5018–5035, 1997.

[32] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, April 1997.

[33] A. Jasra, A. Doucet, D. A. Stephens, and C. C. Holmes. Interacting sequential Monte Carlo samplers for trans-dimensional simulation. *Computational Statistics & Data Analysis*, 52(4):1765–1791, 2008.

[34] A. Jasra, D. A. Stephens, and C. C. Holmes. On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279, 2007.

[35] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3:300, 1935.

[36] P. S. Koutsourelakis. Design of complex systems in the presence of large uncertainties: A statistical approach. *Computer Methods in Applied Mechanics and Engineering*, 197(49-50):4092–4103, 2008.

[37] P. S. Koutsourelakis. Accurate uncertainty quantification using inaccurate models. *SIAM Journal of Scientific Computing*, 31(5):3274–3300, 2009.

[38] S. Kullback. *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.

[39] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[40] H. Künsch. Recursive monte carlo filters: Algorithms and theoretical analysis. *The Annals of Statistics*, (5):19832021, 2005.

[41] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562, 2002.

[42] T. Lelièvre, M. Rousset, and G. Stoltz. Computation of free energy dierences through nonequilibrium stochastic dynamics: The reaction coordinate case. *The Journal of Chemical Physics*, 222(2):624–643, 2007.

[43] T. Lelièvre, M. Rousset, and G. Stoltz. Computation of free energy profiles with parallel adaptive dynamics. *The Journal of Chemical Physics*, 126(13):134111, 2007.

[44] T. Lelièvre, M. Rousset, and G. Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21:1155–1181, 2008.

[45] T. Lelièvre, M. Rousset, and G. Stoltz. *Free Energy Computations: A Mathematical Perspective*. Imperical College Press, 2010.

[46] F. Liang. Generalized Wang-Landau algorithm for Monte Carlo Computation. *J. Amer. Statist. Assoc.*, 100:1311–1327, 2005.

[47] F. Liang, C. Liu, and R. J. Carroll. Stochastic Approximation in Monte Carlo Computation . *J. Amer. Statist. Assoc.*, 102:305–320, 2007.

[48] F. Liang, C. Liu, and R.J. Carroll. *Advanced Markov chain Monte Carlo: Learning from Past Samples*. Wiley., 2010.

[49] J S Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001.

[50] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22:551575, 1977.

[51] L. Maraglian and E. Vanden-Eijnden. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical Physics Letters*, 426(1-3):168–175, July 2006.

[52] L. Maragliano and E. Vanden-Eijnden. Single-sweep methods for free energy calculations. *The Journal of Chemical Physics*, 128(18):184110, May 2008.

[53] S. Marsili, A. Barducci, R. Chelli, P. Procacci, and V. Schettino. Self-healing umbrella sampling: a non-equilibrium approach for quantitative free energy calculations. *The Journal of Physical Chemistry B*, 110(29):14011–14013, July 2006.

[54] F. J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46:68–78, 1951.

[55] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[56] H. Meirovitch. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Current Opinion in Structural Biology*, 17(2):181–186, 2007.

[57] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[58] M. Métivier and P. Priouret. Applications of a Kushner and Clark lemma to a general classes of stochastic algorithms. *IEEE Transactions on Information Theory*, 30:140151, 1984.

[59] R. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

[60] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[61] J. M. Rickman and R. LeSar. Free-energy calculations in materials research. *Annual Review of Materials Research*, 32:195–217, 2002.

[62] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[63] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

[64] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[65] N. Schraudolph and T. Graepel. Towards stochastic conjugate gradient methods. In *Proceedings of the 9th International Conference on Neural Information Processing*, Singapore, 2002.

[66] M.R. Shirts and J.D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129:124105, 2008.

[67] Marc Souaille and Benoit Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1):40 – 57, 2001.

[68] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley and Sons, 2003.

[69] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti. Bond-orientational order in liquids and glasses. *Physical Review B*, 28(2):784–805, 1983.

[70] G. Stoltz. Path sampling with stochastic dynamics: Some new algorithms. *Journal of Computational Physics*, 225:491–508, 2007.

[71] R. Swendsen, M. Fasnacht, and J. Rosenberg. The adaptive integration method for calculating general free energy functions. *Computer Physics Communications*, 169(1-3):274–276, 2005.

[72] M. E. Tipping. The Relevance Vector Machine. In *Advances in Neural Information Processing Systems 12*, pages 652–658. MIT Press, 2000.

[73] M. T. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[74] A. B. Tsybacov. *Introduction to Nonparametric Estimation*. Springer., 2009.

[75] A. F. Voter. A method for accelerating the molecular dynamics simulation of infrequent events. *The Journal of Chemical Physics*, 106(11):4665–4677, 1997.

[76] F. Wang and D. P. Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Physical Review E*, 64(5):56101, 2001.

[77] F. M. Ytreberg, R. H. Swendsen, and D. M. Zuckerman. Comparison of free energy methods for molecular systems. *The Journal of Chemical Physics*, 125(18):184114, 2006.

[78] A.L.. Yuille. The convergence of contrastive divergences. In *Advances in Neural Information Processing Systems 17. NIPS.*, December 2004.

[79] S. C. Zhu, Y. Wu, and D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8):1627–1660, 1997.

[80] S. C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.